

Enhancing Archival Access with Artificial Intelligence (AI)

Briana Giasullo, Academy of Natural Sciences
of Drexel University

Also with:

Chrysanthemum Lovelace, University of Pittsburgh

Carolyn Friedrich, University of Pittsburgh

About the Academy of Natural Sciences

- In addition to the public-facing museum, the Academy has scientists working behind the scenes
- The Center for Systematic Biology and Evolution (CSBE) cares for and studies over 18 million specimens of plants and animals, and conducts research in systematics, ecology, evolution, and paleontology.



Bird specimens from the Ornithology department

About the Academy Library and Archives

- Three main sections: Reference library, Wolf room (rare books), and the archives
- Mainly serve Academy staff
- Public access by appointment
- Participate in Academy-wide events such as Members Night and Bug Fest
- Increasing our partnerships with Drexel students and faculty, research fellows, and local organizations
- Also looking to increase worldwide access to collections using various online platforms



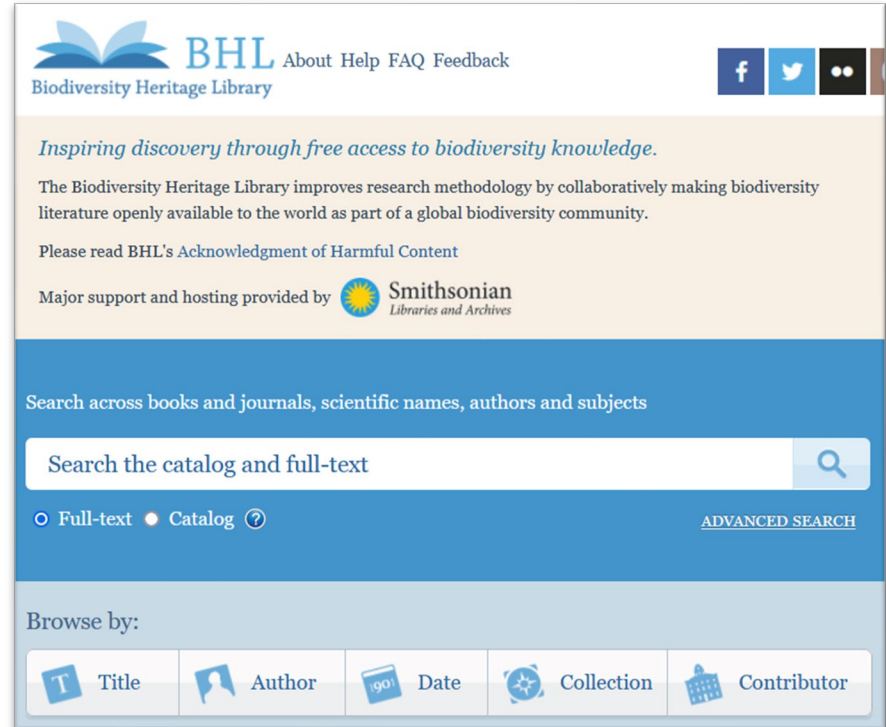
Engravings by Helen E. Lawson (ANSP-Coll-0079)

Background

- Biodiversity Heritage Library (BHL)
- Academy's relationship with BHL
- OCR and HTR
- Field journals
- Project goals

Biodiversity Heritage Library (BHL)

- Free, global access to biodiversity literature
- Currently 643 contributors from around the world providing digital images from their collections
- Anyone can use BHL to view digitized content from natural science organizations
- Great for researching biodiversity data, viewing natural illustrations for art inspiration, and more

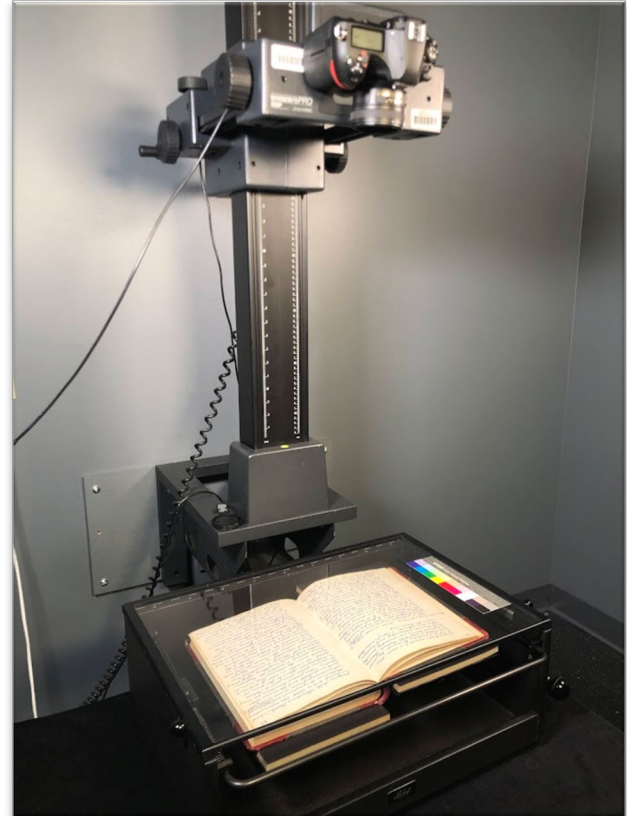


The screenshot shows the BHL website interface. At the top, there is a logo for BHL (Biodiversity Heritage Library) with navigation links for 'About', 'Help', 'FAQ', and 'Feedback'. Social media icons for Facebook, Twitter, and YouTube are also present. Below the logo, a tagline reads: "Inspiring discovery through free access to biodiversity knowledge." A paragraph explains that BHL improves research methodology by collaboratively making biodiversity literature openly available. A link to "Please read BHL's Acknowledgment of Harmful Content" is provided. Below this, it states "Major support and hosting provided by" followed by the Smithsonian Libraries and Archives logo. The main search area features a blue header with the text "Search across books and journals, scientific names, authors and subjects". A search input field contains the placeholder text "Search the catalog and full-text" and a magnifying glass icon. Below the search field, there are radio buttons for "Full-text" (selected) and "Catalog", along with a help icon. An "ADVANCED SEARCH" link is located to the right. At the bottom, a "Browse by:" section offers five options: "Title" (with a 'T' icon), "Author" (with a person icon), "Date" (with a '1901' icon), "Collection" (with a star icon), and "Contributor" (with a building icon).

Screenshot of BHL search box

BHL and the Academy

- Affiliated since 2010
- Digitization by request
- Current process is inconsistent and infrequent
- Need a streamlined, consistent approach (which this project will help with)
- Getting volunteers and students to help with the digitization process



Overhead camera in the Library's Greenfield center

Page 65 (Text)
Page 66 (Text)
Page 67 (Text)
Page 68 (Text)
Page 69 (Text)
Page 70 (Text)
Page 71 (Text)
Text
Illustration, Text
Text
Text
[Show More](#)

URL for Current Page
<https://www.biodiversitylibrary.org/page/40356926>

Scientific Names on this Page

Page 70 (Text)

[Corvus splendens zugmayeri Laubmann, 1913](#) 🔍 ▶
[Falco Linnaeus, 1758](#) 🔍 ▶
[Sarcogrammus indicus](#) 🔍 ▶
[Sarcogrammus indicus aigneri](#) 🔍 ▶
[Sarcogrammus Reichenbach, 1852](#) 🔍 ▶

Indexed by [Global Names](#) 🌐

70

87. *Sarcogrammus indicus aigneri* Laubmann.

Sarcogrammus indicus aigneri Laubmann Falco IX, p. 30 (1913 — Sonmiani).
Sarcogrammus indicus Blanford, Birds Brit. India, Vol. IV, p. 224.

Nr. 41 ♂ Sonmiani, Mekran, 2. III. 1911: a. 235, r. 34 (Typus).
Nr. 354 ♂ Las Bela, 23. III. 1911: a. 215, r. 35.
Nr. 408 ♂ Las Bela, 27. III. 1911: a. 228, r. 33.
Nr. 415 ♂ Las Bela, 29. III. 1911: a. 218, r. 32.

Ich habe bereits gelegentlich der Charakterisierung der neuen Form auf die Unterschiede aufmerksam gemacht, die zwischen den beiden Formen *Sarcogrammus indicus indicus* (Bodd.) und *Sarcogrammus indicus aigneri* Laubm. bestehen und die Veranlassung zur Abtrennung der Balutschistan-Vögel gegeben haben.

Das Allgemeinaussehen der neuen Form ist ein auffallend fahles im Gegensatz zu dem indischen Vogel, bei dem dunklere Tönung vorherrscht. So fehlen bei *Sarcogrammus indicus aigneri* die metallisch grünen Töne in der Färbung der Oberseite fast ganz, ebenso sind die metallisch purpurnen Farben an den Oberflügeldecken äußerst reduziert. Außerdem reicht auch die schwarze Färbung nicht so weit in den Nacken herunter wie bei indischen Exemplaren.

Aber nicht nur hinsichtlich der Färbung des Gefieders bestehen charakteristische Unterschiede zwischen beiden Formen, sondern auch in Bezug auf die Größenverhältnisse ergaben sich Verschiedenheiten. So stellte es sich durch die von mir vorgenommenen Messungen heraus, daß die Balutschistan-Form etwas längere Flügelmaße aufweist und auch etwas längere Schnabeldimensionen besitzt als Stücke von Indien und Ceylon, die mir zum Vergleich zu Gebote standen.

Ich führe hier zur genaueren Orientierung die gefundenen Maße für die 4 Stücke aus Ceylon und Indien an:

Contributed by [American Museum of Natural History Library](#) 🌐

1. ♂ 12. I. 1905, Nord-Ceylon: a. 209, r. 34.
2. ♀ 10. I. 1905, Nord-Ceylon: a. 215, r. 30.
3. ♂ 10. I. 1905, Nord-Ceylon: a. 212, r. 30.
4. Indien, Zentralprovinzen: a. 214, r. 29, defekt.

Uncorrected OCR ?

Nr. 408 5 Las Bela, 27. III. 1911: a. 228, r. 33.
Nr. 415 5 Las Bela, 29. III. 1911: a. 218, r. 32.

Ich habe bereits gelegentlich der Charakterisierung der neuen Form auf die Unterschiede aufmerksam gemacht, die zwischen den beiden Formen *Sarcogrammus indicus indicus* (Bodd.) und *Sarcogrammus indicus aigneri* Laubm. bestehen und die Veranlassung zur Abtrennung der Balutschistan-Vögel gegeben haben.

Das Allgemeinaussehen der neuen Form ist ein auffallend fahles im Gegensatz zu dem indischen Vogel, bei dem dunklere Tönung vorherrscht. So fehlen bei *Sarcogrammus indicus aigneri* die metallisch grünen Töne in der Färbung der Oberseite fast ganz, ebenso sind die metallisch purpurnen Farben an den Oberflügeldecken äußerst reduziert. Außerdem reicht auch die schwarze Färbung nicht so weit in den Nacken herunter wie bei indischen Exemplaren.

Aber nicht nur hinsichtlich der Färbung des Gefieders bestehen charakteristische Unterschiede zwischen beiden Formen, sondern auch in Bezug auf die Größenverhältnisse ergaben sich Verschiedenheiten. So stellte es sich durch die von mir vorgenommenen

OCR on Handwriting

Pages

- Page [26] (Text)
- Page [27] (Text)
- Page [28] (Text)
- Page [29] (Blank)
- Page [30] (Text)
- Page [31] (Text)
- Page [32] (Text)
- Page [33] (Text)
- Page [34] (Text)
- Page [35] (Text)
- Page [36] (Text)

Show More

URL for Current Page

<https://www.biodiversitylibrary.org/page/59317284>

Scientific Names on this Page

Page [33] (Text)

No Scientific Names found

78. alt. 6/48 ft. IX. 13. 1907

Williams Arizona. All specimens taken in open place thickly overgrown with rabbit weed and other equally low green herbs - a country of many glades where all the ground is green and the scattered pieces everywhere produce a park-like effect. But

Contributed by Academy of Natural Sciences of Drexel University, Library and Archives

inside info text

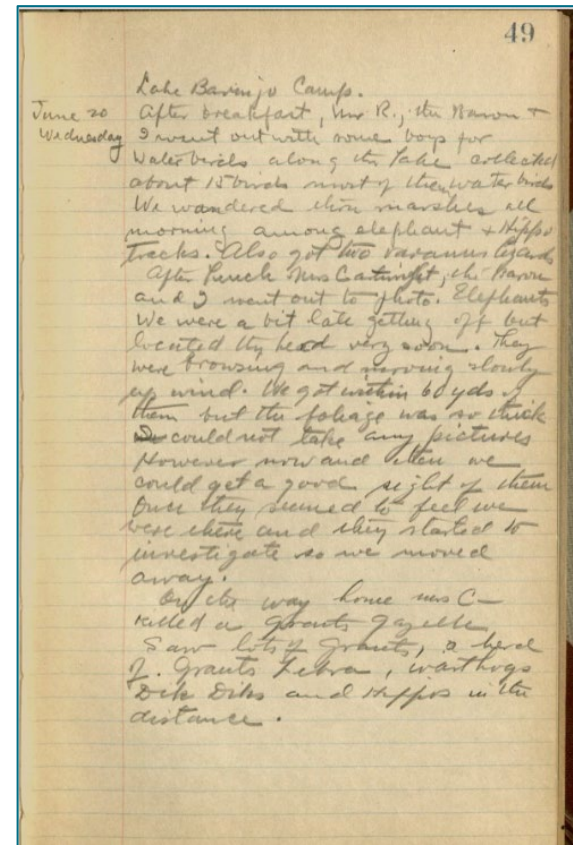
Uncorrected OCR

```
fu
- ~cj/iC(L&6y L&-*L)
WCaV* ^ ~£S*vc^tiZ7
■ J 7 ^CudLzx : ^jrl -' -
T - ~£s lif-u*4jl A I /j m
1 P&LC.. /'J). Otz/Zz' i C($ 1
A/eu *j tiJ J/tjp # 7/ ^<i/ I - /
j <v,
,3H
■ - ^CZ"Gi,t. f. £X
■ ii r y y ^, W** -
I' H / <= tof* e &%ah. i. - £u
I .... // AU-e^TZ [ _
m-s, **Thi%' f / - ^ c' c ^ /
i-^T
i-
M& I' /S <J Os?,*^ /f/ y « L. d //r
Ly/// ^ ir^fy^
j/j » ,r -* . <-fr* £Cuy ., x/ K
4 t £ / /f 4 C Vv /' A/. 4^ , /icv <* /74
■ U lunS/;- X.
J./Z tjW'7&, j% \
. ttfiSf&t ct^y
• ^■&H. Si edL^ ?7cuV
r 1
u^ . ^ . c>
, Q/Cts^
9
```

Field journal of J.A.G. Rehn kept during a collecting trip to Arizona, California, Nevada, New Mexico, and Texas, 1907

Field Journals

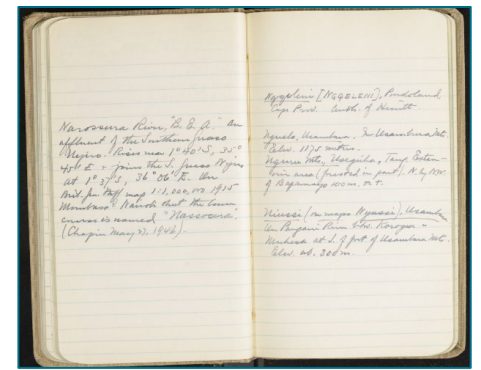
- Written by naturalists while they are out collecting specimens and conducting research
- Contain valuable biodiversity data about specimens collected, environmental attributes, weather, and more
- Sometimes contain narratives about the daily life of the author, including details about travel, meals, and general thoughts
- The Academy holds many field journals in its archives, only some of which are digitized (even fewer available online)
- Handwritten, often in cursive, sometimes using symbols or shorthand, making traditional optical character recognition (OCR) difficult or impossible



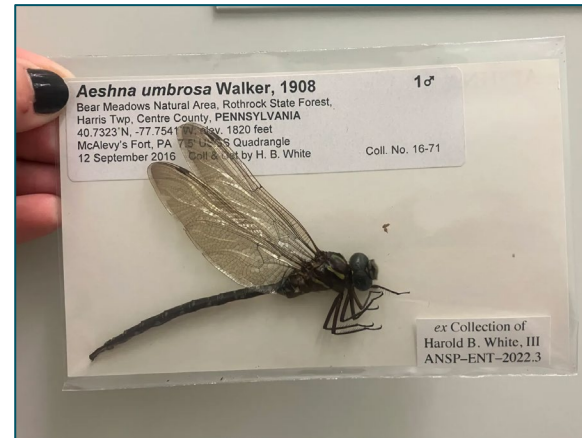
Field journal in handwritten script, possibly by Harold Tichenor Green, 1934 (ANSP-Coll-0998)

Hence the Project

- Help BHL by adding transcriptions to their database
- Immediately text searchable to all BHL users
- With text-searchable field journals, we are better able to connect the journals to specimens in other Academy departments (e.g. a specimen in Entomology can be connected to the journal written by the naturalist who collected it)
- We can also think about specimens in new ways using the **extended specimen** concept
 - Example: Specimens as culturally situated objects, with context and history surrounding their collection



List of wet specimens written by James A.G. Rehn, 1934



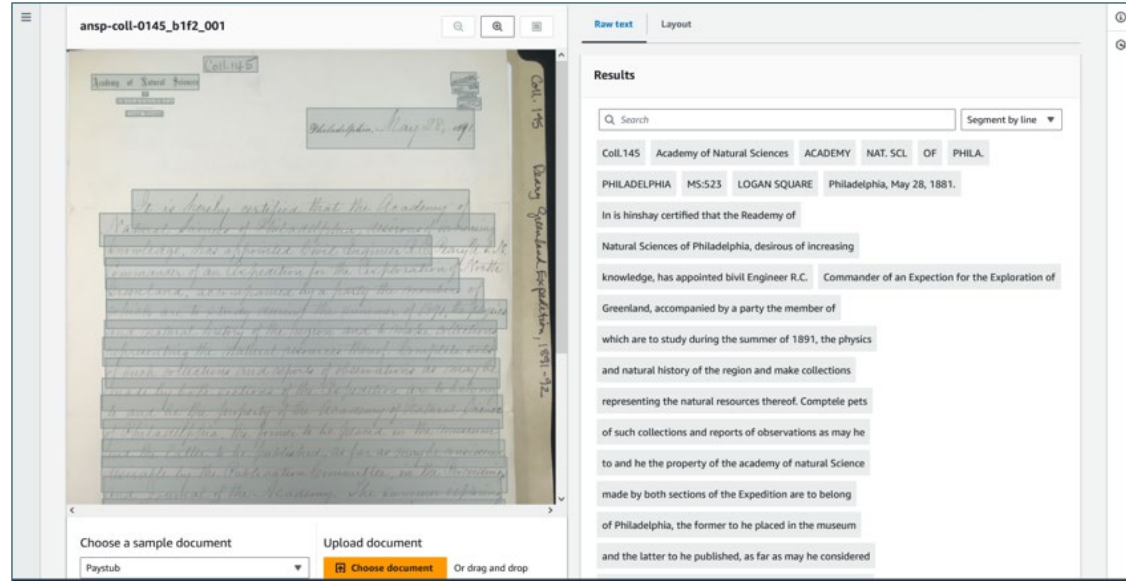
Specimen from the Entomology department

The Actual Work

- Amazon Textract
- Zooniverse
- Overview of steps
- The final product!
- Status
- Lessons learned

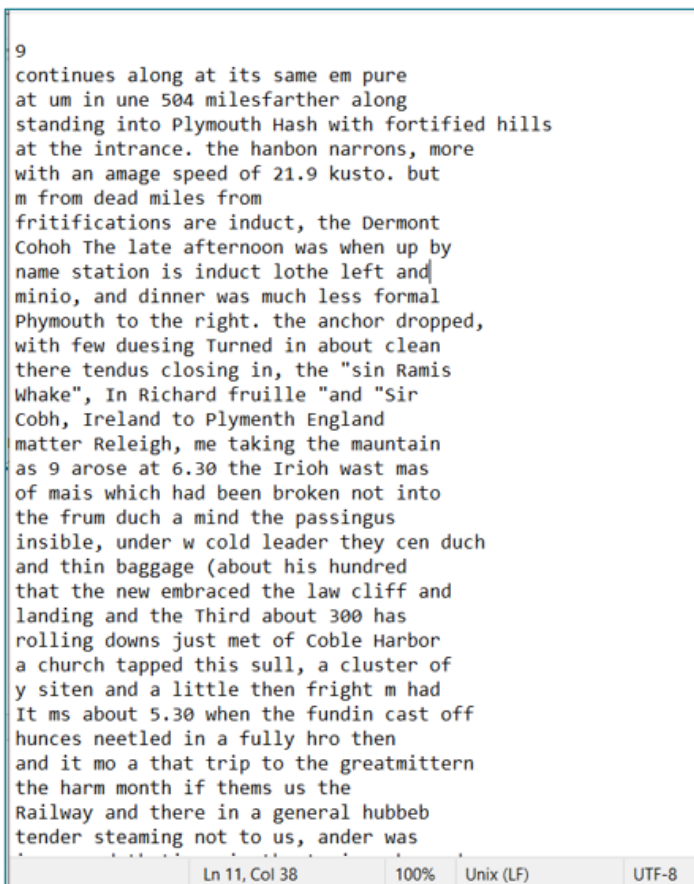
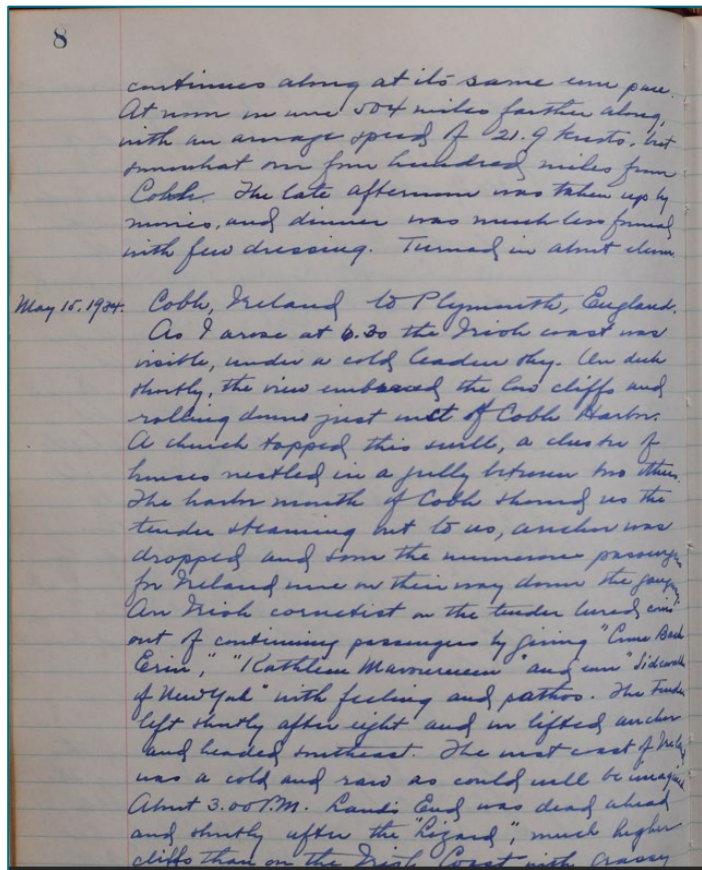
Amazon Textract

- Machine learning (ML) service that automatically extracts text, handwriting, layout elements, and data from scanned documents.
- More than optical character recognition (OCR)
- Can be trained to read handwriting and extract specific data from documents
- Selected due to the Academy's existing relationship with Amazon as a cloud storage solution
- Cost is based on use (so relatively cheap for us)



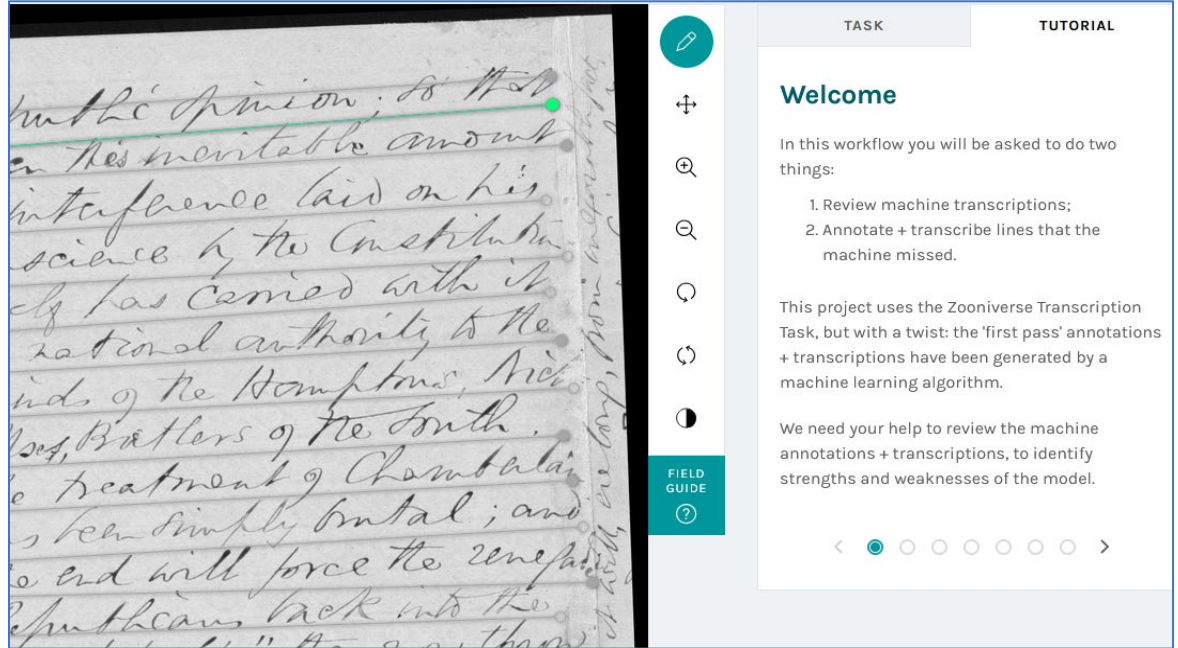
Textract free-trial demo

Textextract sample image and text



Zooniverse

- Crowdsourcing platform
- Can create a project, write tutorials, and recruit volunteers to help with tasks
- Often used for identifying and tagging things in photographs (species, planets, etc.)
- Great for transcription projects, specifically with correcting AI with handwritten documents that elude traditional OCR!



The screenshot displays a Zooniverse transcription task interface. On the left, a handwritten document is shown with a green line highlighting a portion of the text. The text is written in cursive and includes phrases like "public opinion; as the", "in this inevitable amount", "interference laid on his", "science by the Constitution", "ly has carried with it", "rational authority to the", "inds of the Democrats, Nicks", "sons, Brattlers of the South.", "e treatment of Chamberlain", "s been simply brutal; and", "e end will force the renegade", "Republicans back into the", "it will be long, from inter-".

On the right, a tutorial panel is visible with the following content:

TASK **TUTORIAL**

Welcome

In this workflow you will be asked to do two things:

1. Review machine transcriptions;
2. Annotate + transcribe lines that the machine missed.

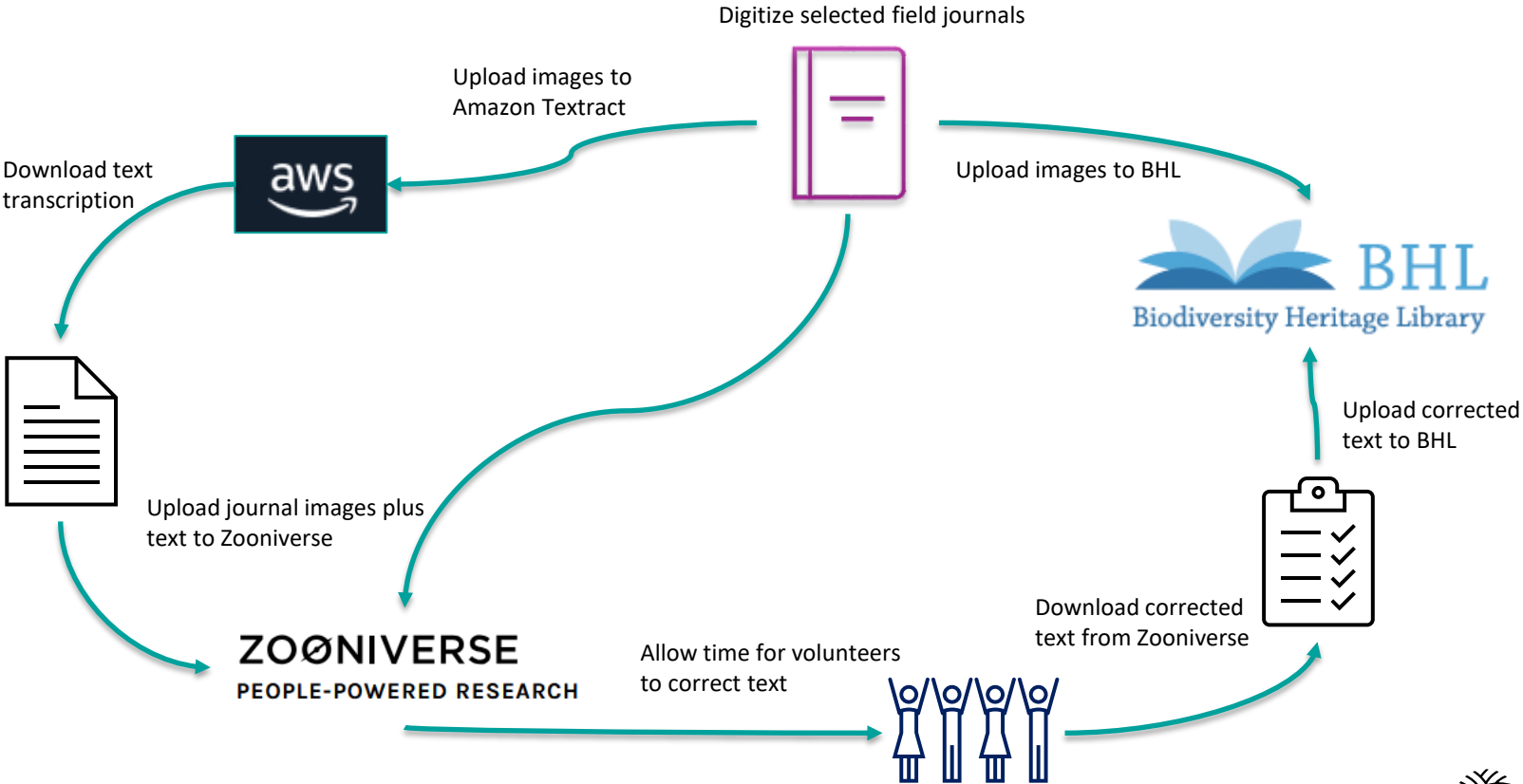
This project uses the Zooniverse Transcription Task, but with a twist: the 'first pass' annotations + transcriptions have been generated by a machine learning algorithm.

We need your help to review the machine annotations + transcriptions, to identify strengths and weaknesses of the model.

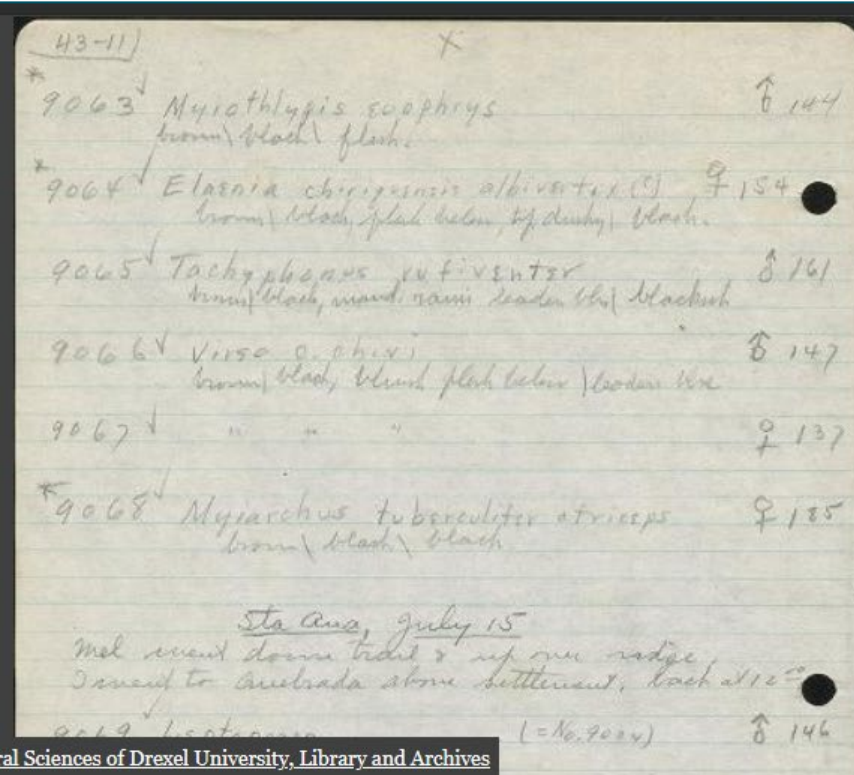
Navigation icons include a pencil, a plus sign, a magnifying glass, a search icon, a refresh icon, a back icon, a forward icon, and a field guide icon (a question mark in a circle).

Sample transcription project on Zooniverse

Breakdown of Steps



Et voila!



al Sciences of Drexel University, Library and Archives

Manual Transcription ?

43-11

*9063 ✓ *Myiothlypis euophrys* [male] 144

brown\ black\ flesh.

*9064 ✓ *Elaenia chiriquensis albivertex* (?)
[female] 154

brown\ black, flesh below, tip dusky\ black

9065 ✓ *Tachyphonus rufiventer* [male] 161

brown\ black, mand. rami leaden blue\ blackish

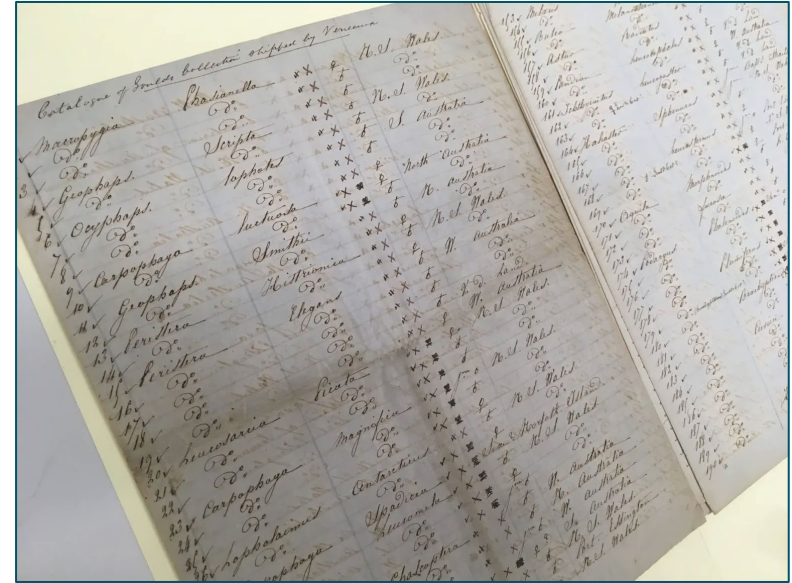
9066 ✓ *Vireo c. chivi* [male] 147

This transcription was done manually by human volunteers, not by machine, but shows what completed transcriptions could look like.

Melbourne Armstrong Carriker, 1st Bolivian Expedition (numbers 9000-11269)

Project Status (as of November 2024)

- Several field journals have been digitized
- National Endowment for the Humanities (NEH) grant submitted (decision expected in December)
 - Primarily will provide funding to hire a temporary employee to assist with digitization, metadata creation, and other tasks
 - Backup: Work study positions
- Amazon Textract has been set up
- Zooniverse project is in the works
- Beginning to draft documentation for the process so the work can be sustainable in the future



Verreaux shipping manifest for Gould collection, 1847, Ornithology Department Records, Coll 54.

LESSONS LEARNED! (Possibly most important slide)

- If you can, befriend your IT department!
 - Be honest with them about what you need help with vs. what you can figure out yourself
 - Comply with institutional policies
 - When things break, they will be happier to help you if you collaborated with them early!
- Try a bunch of options as far as software programs and other tools (free trials are great)
- Talk to others doing similar work!
- Stay open-minded
 - I initially was skeptical of AI, integrating platforms, and working with Amazon, but very glad I didn't immediately rule things out!
 - Things change quickly



From the TV show *The IT Crowd*

What else would you like to know?

- I'm always happy to talk more about this project!
- Briana Giasullo: bg557@drexel.edu
- Academy Library and Archives: library@ansp.org
- Research appointments available Tuesday – Friday, 10 AM – 4 PM
- We are always happy to answer questions about our collections and work!

