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

The Academy of Natural Sciences of DREXEL UNIVERSITY

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

The Academy of Natural Sciences of DREXEL UNIVERSITY

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)

The Academy of Natural Sciences of DREXEL UNIVERSITY

Background

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

Biodiversity Heritage Library (BHL)

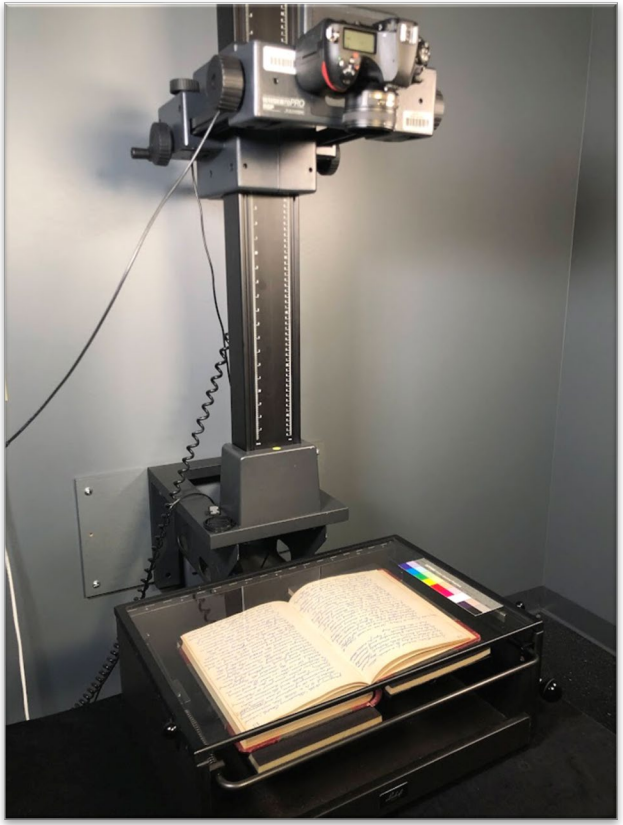
- Free, global access to biodiversity literature
- Currently 643 contributors from around the world providing digital images from their collections
- Anyone can use BHL to view digitized content from natural science organizations
- Great for researching biodiversity data, viewing natural illustrations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

Screenshot of BHL search box

The Academy of Natural Sciences of DREXEL UNIVERSITY

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

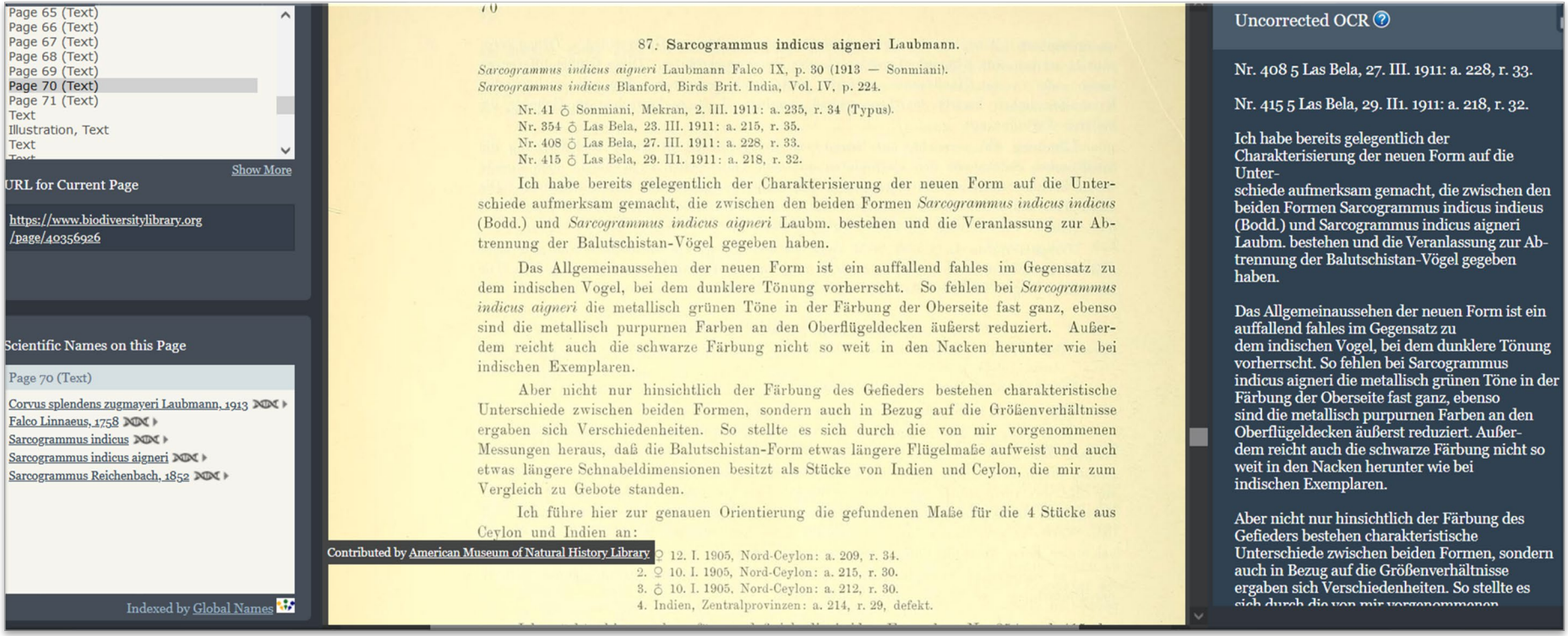


Overhead camera in the Library's Greenfield center

The Academy of Natural Sciences of DREXEL UNIVERSITY

OCR in BHL

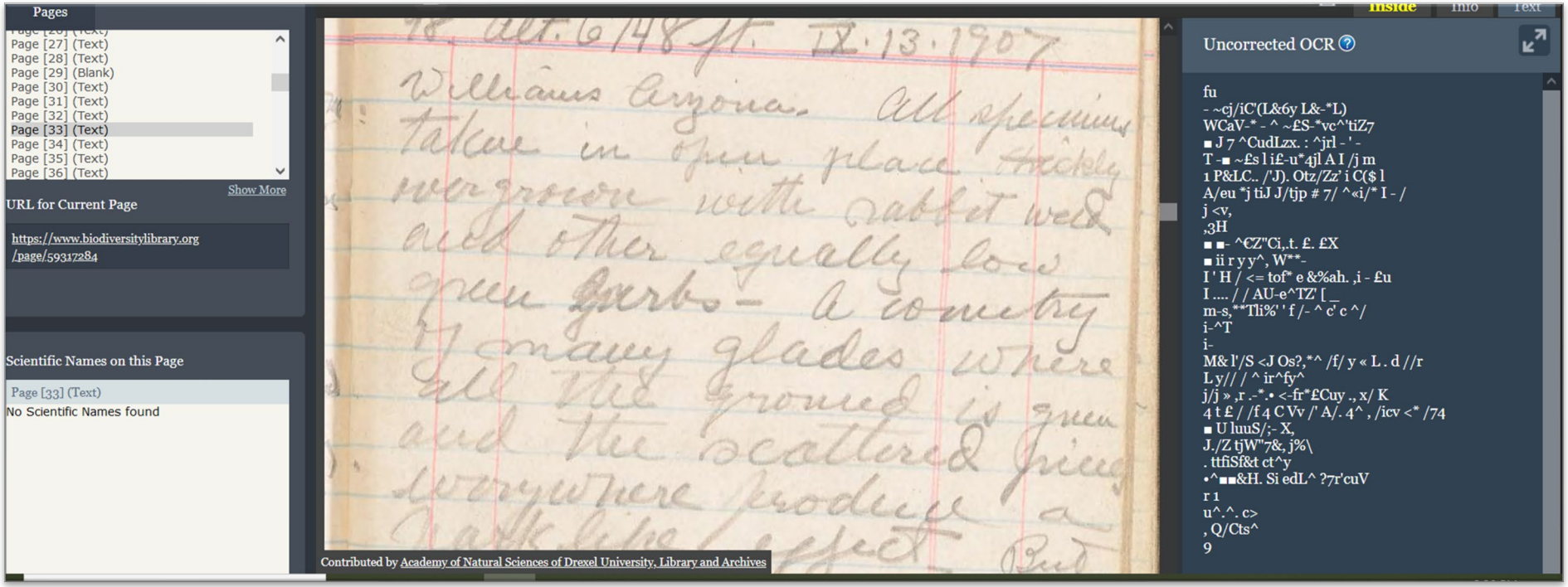
BHL content automatically goes through optical character recognition (OCR) upon upload, making everything text-searchable and allowing for taxonomic name recognition



Abhandlungen der Mathematisch-Physikalischen Klasse der Königlich Bayerischen Akademie der Wissenschaften

The Academy of Natural Sciences of DREXEL UNIVERSITY

OCR on Handwriting

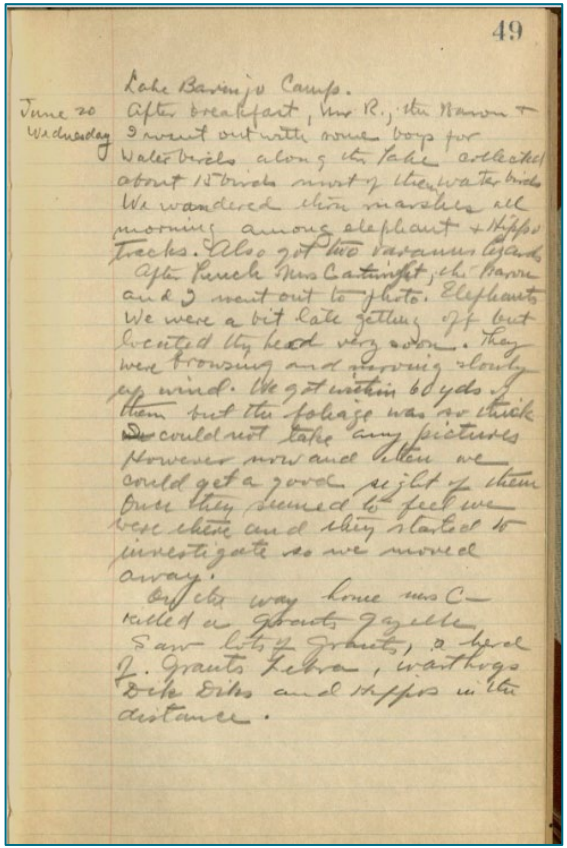


Field journal of J.A.G. Rehn kept during a collecting trip to Arizona, California, Nevada, New Mexico, and Texas, 1907

The Academy of Natural Sciences of DREXEL UNIVERSITY

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


49

Lake Baringo Camp.

June 20 Wednesday After breakfast, Mrs R., the Nanyuki and I went out with some boys for Water birds along the lake collected about 15 birds most of them water birds We wandered thru marshes all morning among elephant + hippo tracks. Also got two varanus lizards

After lunch Mrs C. retired, the Baron and I went out to photo. Elephants We were a bit late getting off but I crawled thru head very soon. They were browsing and moving slowly up wind. We got within 60 yds of them but the foliage was so thick I could not take any pictures However now and then we could get a good sight of them but they sensed us & fled we were there and they started to investigate so we moved away.

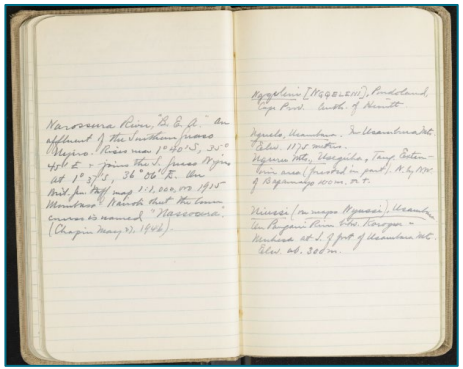
On the way home Mrs C— killed a Grants Gazelle Saw lots of Grants, 2 herds of Grants Zebra, warthogs Dik Diks and hippos in the distance.

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

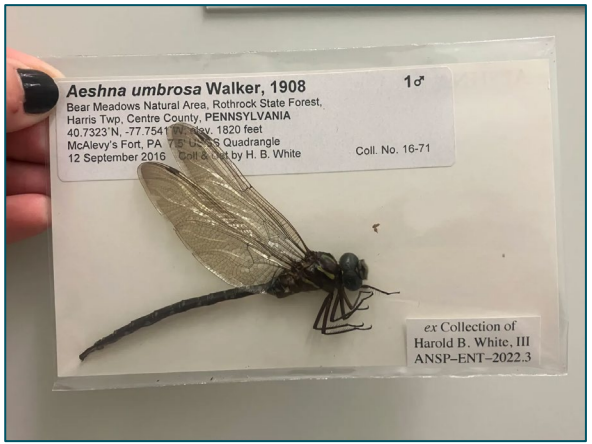
The Academy of Natural Sciences of DREXEL UNIVERSITY

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

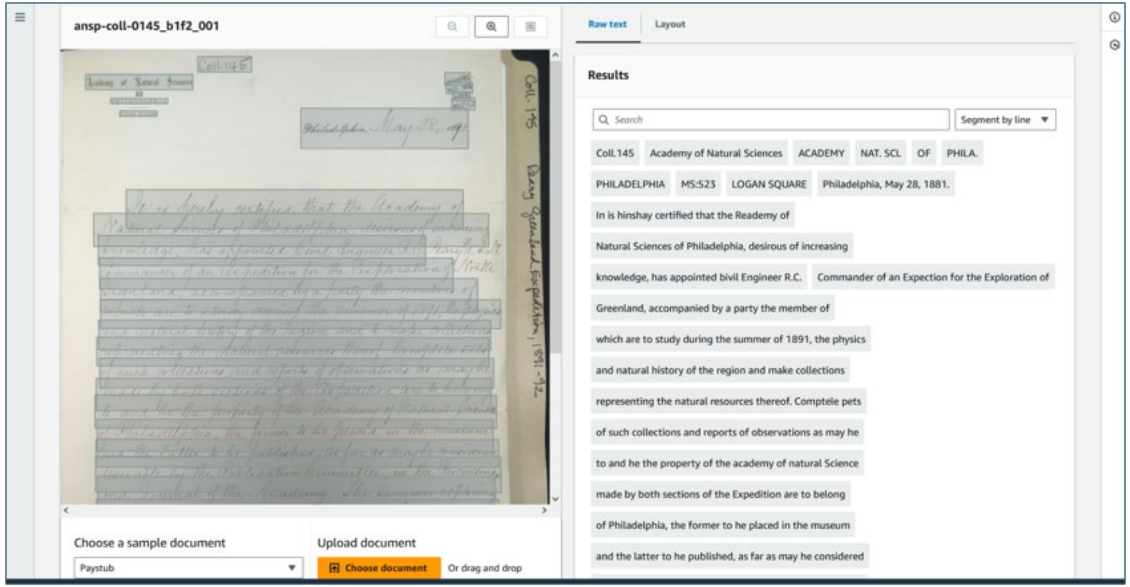
The Academy of Natural Sciences of DREXEL UNIVERSITY

The Actual Work

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

Amazon Textract

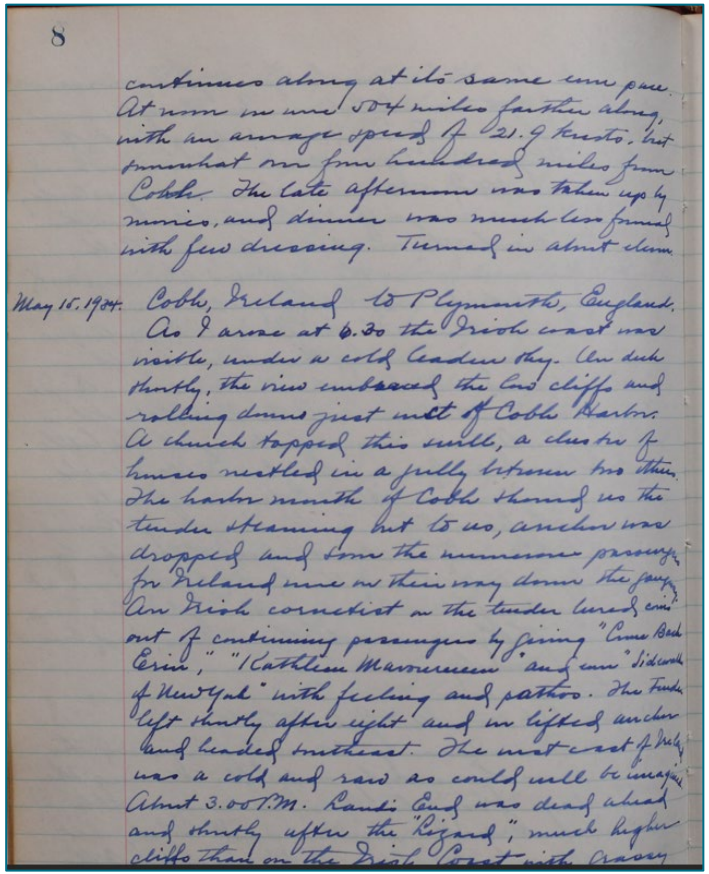
- Machine learning (ML) service that automatically extracts text, handwriting, layout elements, and data from scanned documents.
- More than optical character recognition (OCR)
- Can be trained to read handwriting and extract specific data from documents
- Selected due to the Academy's existing relationship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



Textract free-trial demo

The Academy of Natural Sciences of DREXEL UNIVERSITY

Textract sample image and text



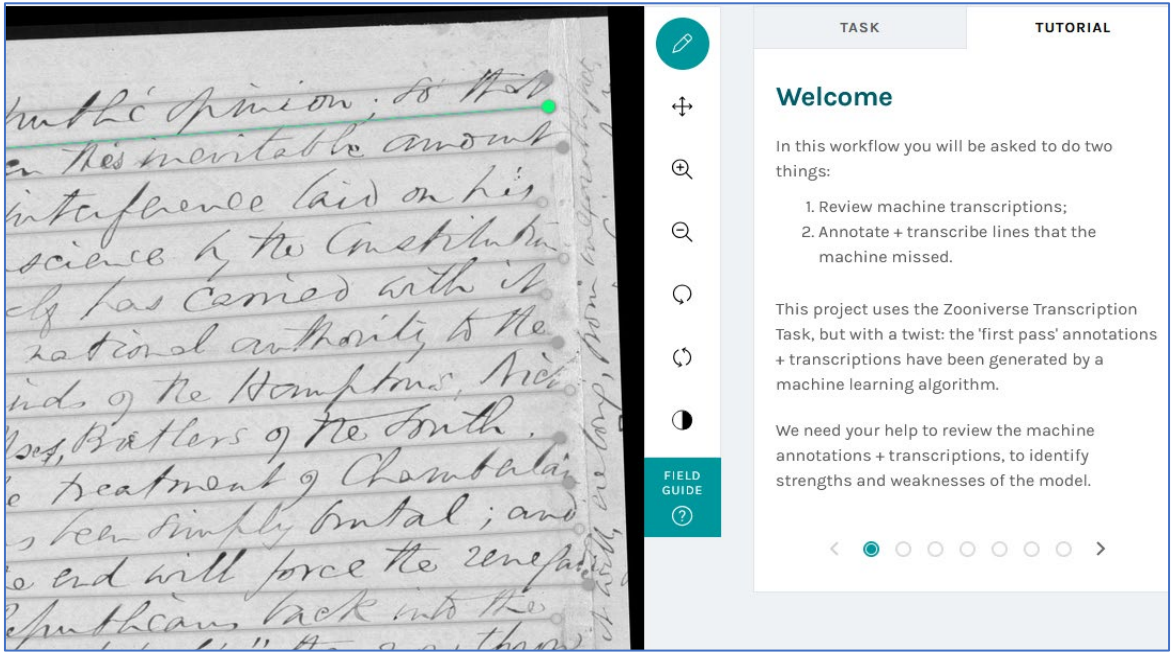

```
9
continues along at its same em pure
at um in une 504 milesfarther along
standing into Plymouth Hash with fortified hills
at the intrance. the hanbon narrons, more
with an amage speed of 21.9 kusto. but
m from dead miles from
fritifications are induct, the Dermont
Cohoh The late afternoon was when up by
name station is induct lothe left and
minio, and dinner was much less formal
Phymouth to the right. the anchor dropped,
with few duesing Turned in about clean
there tendus closing in, the "sin Ramis
Whake", In Richard fruille "and "Sir
Cobh, Ireland to Plymenth England
matter Releigh, me taking the mauntain
as 9 arose at 6.30 the Irioh wast mas
of mais which had been broken not into
the frum duch a mind the passingus
insible, under w cold leader they cen duch
and thin baggage (about his hundred
that the new embraced the law cliff and
landing and the Third about 300 has
rolling downs just met of Coble Harbor
a church tapped this sull, a cluster of
y siten and a little then fright m had
It ms about 5.30 when the fundin cast off
hunces neetled in a fully hro then
and it mo a that trip to the greatmittern
the harm month if thems us the
Railway and there in a general hubbeb
tender steaming not to us, ander was
```

Ln 11, Col 38 | 100% | Unix (LF) | UTF-8

The Academy of Natural Sciences of Drexel University

Zooniverse

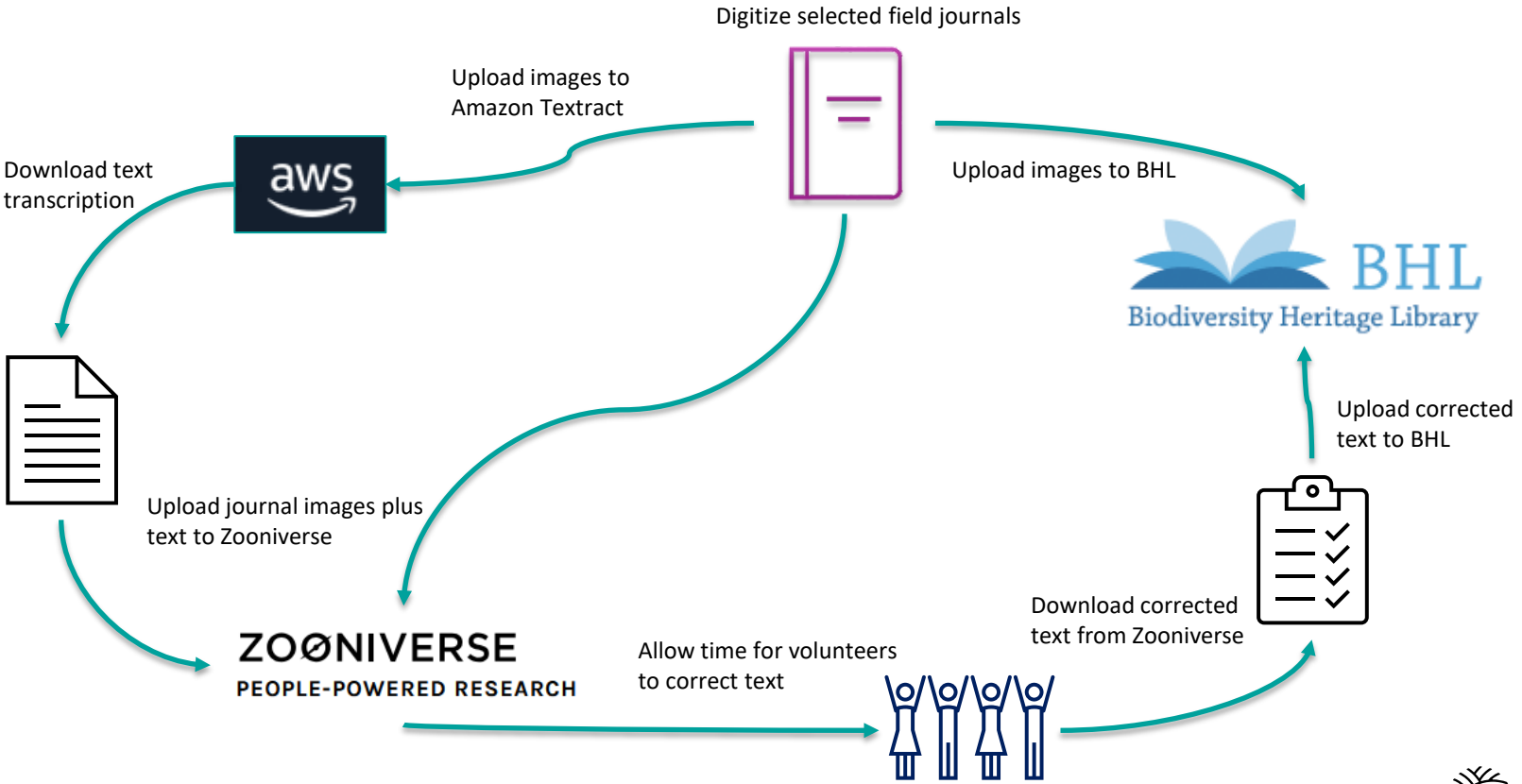
- Crowdsourcing platform
- Can create a project, write tutorials, and recruit volunteers to help with tasks
- Often used for identifying and tagging things in photographs (species, planets, etc.)
- Great for transcription projects, specifically with correcting AI with handwritten documents that elude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



Sample transcription project on Zooniverse

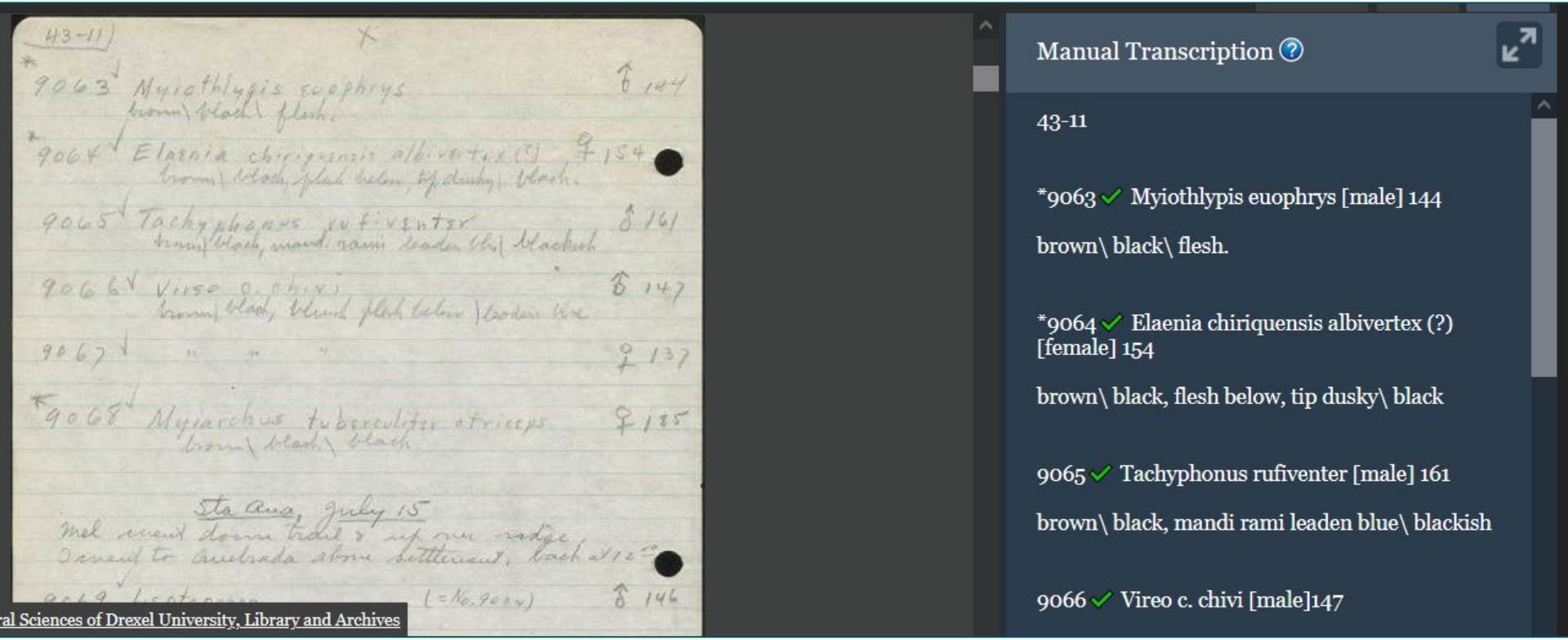
The Academy of Natural Sciences of DREXEL UNIVERSITY

Breakdown of Steps



The Academy of Natural Sciences of DREXEL UNIVERSITY

Et voila!



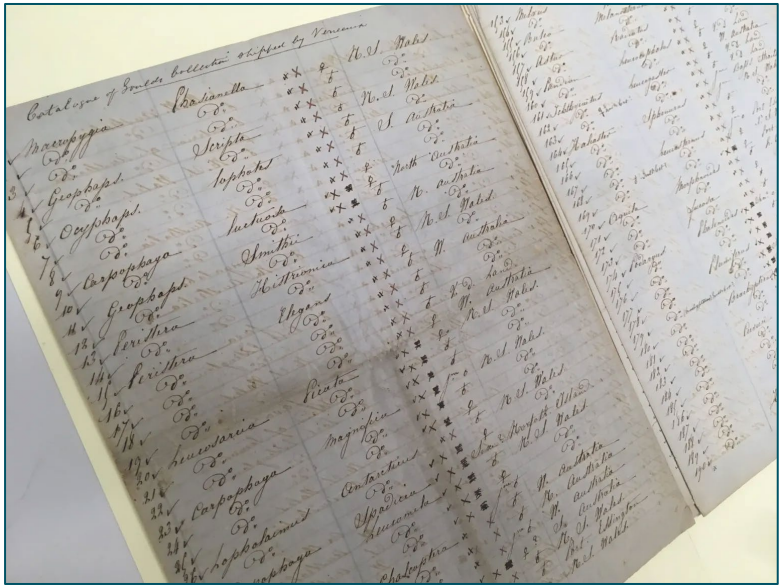
This transcription was done manually by human volunteers, not by machine, but shows what completed transcriptions could look like.

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

The Academy of Natural Sciences of DREXEL UNIVERSITY

Project Status (as of November 2024)

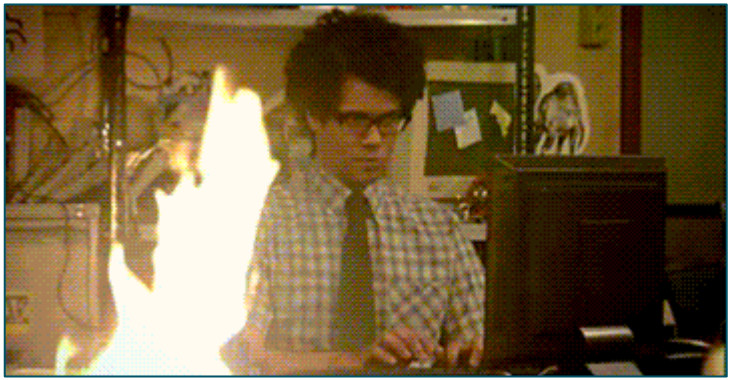
- Several field journals have been digitized
- National Endowment for the Humanities (NEH) grant submitted (decision expected in December)
  - Primarily will provide funding to hire a temporary employee to assist with digitization, metadata creation, and other tasks
  - Backup: Work study positions
- 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!

The Academy of Natural Sciences of DREXEL UNIVERSITY