Webinar & Demo: Strategies for an OCR directed workflow.

When: August 25th, 11 AM EDT
Where: http://idigbio.adobeconnect.com/augmentocr (headsets recommended)
Presenter: Stephen Gottschalk, NYBG Project Coordinator

Links of possible interest:

- aOCR Working Group Wiki: https://www.idigbio.org/wiki/index.php/Augmenting_OCR
  - Join us - Ask us!
- OCR resources: https://www.idigbio.org/wiki/index.php/OCR_Resources
  - link to this recording will be on this page
- Workflows with OCR: https://www.idigbio.org/wiki/index.php/OCR_/_NLP_Workflows
- DRAFT: aOCR + DROID Workflow to incorporate OCR into a digitization workflow: https://docs.google.com/document/d/1O1joFj-jKpWMydIEOtud_7ne5nAsXAPwrino1S6os5Y/edit
- iDigBio Webinar: Visualize Your Text Data Using OCR Output https://www.idigbio.org/content/idigbio-webinar-visualize-your-text-data-using-ocr-output
- Recent publication by an aOCR wg member: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4086207/

Introductions and Recording.

Stephen's presentation and the recording will be available at iDigBio. Here:
OCR resources: https://www.idigbio.org/wiki/index.php/OCR_Resources

Your questions / insights here!

What are you hoping to learn here? What are your questions?

the challenge of skeletal records

how to attach them to field books?
how to not duplicate effort
want themed sets of records for crowd-sourcing
and a method to send more difficult labels to certain individuals

saving time by grouping labels and digitizing a series

how to algorithmically calculate the odds from ocr output to populate certain fields

MySQL Workbench, but should work in any sql database

5000 records - to evaluate takes 25 - 30 minutes. So, seems it would scale into a nightly update.

Brian: So the known are labels that have been visually checked against the OCR? Yes.

Brian: Any paper(s) to be published on this? Yes (from Stephen).

ann molineux: This suggests that you need to run this protocol for every field that may be represented in the label?

From Deb: yes, although you could tailor it. Research (Drinkwater, Cubey, Haston, 2014) Sorting labels by country and collector seems to be what transcribers prefer http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4086207/

Order the labels -- based on the scores -- to take advantage of ditto and human learning!

Access 2007 database > but would be great to put in Symbiota or other databases

Deb: can user make use of / note the failures to some benefit in the workflow? How to ignore Determinations / or use the fails?

Brian: How about a 'confirmed' field next to the various fields so the reviewers can confirm the fields they check against the scan image? Seems like that might be helpful as a person navigates through the records and sorts on different parts of the record (author/country/etc.) Or would that be redundant?

Deb: I was thinking that human input could be worked back into future scoring.

Ann M. asking about Handwriting?

Ann M. What OCR do you use? Software used:

ABBYY FineReader Corporate Edition

setting: maintain line breaks

MySQL Workbench

ACCESS 2007

Python

Thanks everyone! If you've got more questions or requests - please let us know.

Stephen's email: stephen.gottschalk@gmail.com

Stephen will be putting code for this up for everyone (GitHub or DropBox)...

Best!
Deb