Natural Language Processing (NLP) and Machine Learning (ML): Speeding Up Digitization and Increasing User Happiness

OCR fragments – insights and use

Searching OCR text to empower transcription and transcriptionists and scale-up digitization

- Create Your Own Expedition or Record Set Link
- Choose a Country and State/Province
- Select a Symbiota Catalog Number
- Enter a Family Name
- Check all Collections
- OCR Fragments
- Specify search criteria, e.g., Scientific Name
- Match records
- Align OCR
- Generate records

NEW WORK in progress at SYMBOIJA

A batch NLP process tool allowing a collection manager to batch parse stored OCR blocks for 100s of records at a time. Depending on how well OCR parses works for a given collection, processors will have to option to parse only selected targeted fields (e.g. collector, number, date). An option will be available to augment parsed data with content harvested from duplicate records already processed within other institutions.

NEW WORK in progress at NN

NN is collaborating with BiospeX (an iDigBio project), to mine OCR output for descriptive metadata about expedition datasets. FUN, faster, and motivating for users!

General Symbiota Workflow with NLP and ML

A. Images are batch loaded and linked to a new blank record that is only populated with the catalog number obtained from the image file name.
B. The new record is augmented with skeletal data that was obtained during the image process and uploaded as a CSV file. Skeletal data typically consists of fields by scientific name, country, and state, but may include collector, number, date, etc.
C. Automated batch OCR extracts text from the images and stores it in the database linked to the image.
   A. ABBYY OCR engine is typically preferred, but Tesseract can work depending on the font and condition of the text.
D. The data processor uses the OCR parsers integrated into the data entry form to extract content and insert it into the proper Dwc fields.
   A. The lichen and bryophyte portals use the Lbcc parser.
   B. The plant portals use the SALIX parser.
E. Both the LBCC and SALIX parsers are distributed with the Symbiota software and can be activated via the Symbiota configuration file.
   A. The SALIX parser is unique in that it uses word frequency tables to determine which fields, the label content belongs. As specimen records are processed, the frequency tables are augmented, which results in improved label parsing.

Across the planet, groups are working on ways to redesign and optimize current digitization methods – with the goal of achieving industrial scale rates. International collaboration is making it possible to discover and coordinate sharing of skills, software, and emerging new methods.

Worldwide Examples

9 institutes: BGBM, MNHN, MRAC, NHM, NMP, RBGE & RBGK reviewing existing OCR software trials, like those by the iDigBio AOCR working group, and studying automatic metadata tools and workflows. Together they’ve designed an additional trial comparing 6 OCR software programs.

- OCR processing on images provided by 6 partner institutes as well as a set from several US institutes supplied by iDigBio.
- Images include plants, insects, molluscs, and fossils.
- Best OCR results so far:
  - A server-based option (ABBYY Recognition Server v3)
  - A PC option (ABBYY FineReader v12 Professional).
  - Two online service options (Onlineocr.net and Newocr.com) were the best of the online services but did not perform as well as the ABBYY software.
- OCR 100% correct compared to hand transcriptions in some cases.

Via contacts made through the AOCR WG, the SYNTHESYS3 institutes have plans in place to test the LBCC and SERNEC Symbiota Portals which have ML and NLP incorporated in their workflows.

Is Digitization Using These Methods Quicker?

The Royal Botanic Garden Edinburgh (RBGE) used OCR output text to speed up transcribing over 100,000 specimen labels. Creating recordsets for digitization faceted by collector and country averaged 20 minutes (8.9%) faster to digitize per batch of 50 records than the next most efficient method. See: Drinkwater RE, Cubey RWN, Haston EM (2014) The use of Optical Character Recognition (OCR) in the digitisation of herbarium specimen labels. PhytoKeys 38: 15–30. doi: 10.3897/phytokeys.38.7168

Insights shared from NYBG

- Volunteers at NYBG use the LBCC and TTD-TCN Symbiota Volunteer Portals.
- Volunteers use OCR text for copying and pasting long localities.
- OCR fragments make it possible to create suitable records sets designed to maximize transcription efficiency.
- Volunteers use the LBCC Parser in the Bryophyte Crowdsourcing Portal all the time -- it is nearly accurate every time.
- For older and handwritten labels (CINC), transcribers will learn to search records by OCR fragment to create cogent record sets.
- ...and yes, there are challenges: slow servers, the need for more parsers, etc.

What about handwriting?

Ask us about ...
- Transcriptorium and Transcribus
- Berlin Botanic Garden (BGBM)
- Leiden and Naturalis, et al

Got ideas for making digitization faster? Let's talk.