Automatic Metadata Extraction (Darwin Core) From Museum Specimen Labels

P. Bryan Heidorn, Qin Wei
University of Illinois at Urbana-Champaign
The problem

- >1 Billion Natural History Specimens
- Collected over 250 years / many languages
- No publishing standards
- Near infinite classes
  - Your high school teacher lied
- 6 min / label * 1B labels = 100M hours
- Saving 1 min = 16.7 Million hours
- $10/hr = $167,000,000
- 1/4790 of U.S. deregulation financial bailout
Why care

• Historic distribution of species
• Ecological niche modeling (invasiveness, crop hardiness, pest potential)
• Projections of the impact of climate change
• Where did explorer go? (for error detection)
• Will I see a Kirkland Warbler here?
• Do tamaracks grow in sand?
• When did Linden trees bloom before the industrial revolution?
A real-life example: *Baronia brevicornis* and its single food plant, *Acacia cochliacantha* (Soberon)
B. brevicornis Abiotic Niche using BS Garp
II: Estimating the “Area of Accessibility” (Soberon)

• From where? What is the initial condition?
• At what scale? In relation to what vagility parameters?
• At certain scales, one can assume that biogeography is a good surrogate for the accessibility areas, this is, we assume that if a species is present in a given biogeographical region, it can reach all of it.
Natural History Specimens
Sample OCR Output

Yale University Herbarium

YU.001300

Curtisb, North American Pl

C^o.nr r^~n

ANTS,

No. 503* "^ 

Polygala ambiguа, Nutt., var.

Coral soil, Cudjoe Key, South Florida.

Legit A. H. Curtiss.
Label Labels

• bc - barcode
• bt - barcode text
• cm - common/colloquial name
• cn - collection number
• co - collector
• cd - collection date
• fm - family name
• ft - footer info
Label Labels

- gn - genus name
- hd - header info
- in - infra name
- ina - infra name author
- lc - location
- pd - plant description
- sa - scientific name author
- sp - species name
Example Training Record

<?xml version="1.0" encoding="UTF-8"?>
<labeldata>
  <bt>Yale University Herbarium</bt>
  <ns> YU.001300</ns>
  <co cc="Curtiss"> Curtisb, </co>
  <hdlc cc="North American Plants"> North American PI </hdlc>
  <ns>C^o.nr r^-n ANTS,$</ns>
  <cnl> No.$</cnl><cn> 503*$</cn><ns> "^</ns>
  <gn> Polygala</gn><sp> ambigna,$</sp><sa> Nntt.,</sa><val> var.$</val>
  <hb> Coral soil,$</hb><lc> Cudjoe Key, South Florida.</lc>
  <lc><col> Legit</col><co> A. H. Curtiss.$</co>
</labeldata>
Supervised Learning Framework

Training Phase
- Unclassified Labels
  - Human Editing
  - Gold Classified Labels
  - Machine Learner
  - Trained Model

Application Phase
- Unclassified Labels
  - Segmentation
  - Segmented Text
  - Machine Classifier
  - Silver Classified Labels
Herbis Experimental Data

- 295 marked up records
- 74 label states
- 5-fold cross-validation
Performances of NB and HMM

![Graph showing the comparison of NB and HMM performances across different elements.](image)

**Elements**

- bc
- bt
- cd
- cdl
- cm
- cml
- cn
- cnl
- co
- col
- ct
- dtl
- fm
- fml
- gn
- hb
- hbl
- hdlc
- in
- latlon
- lc
- lcl
- pd
- sa
- snl
- sp

**F-Score**

- NB
- HMM
Polygala ambiguа, Nutt., var.

Coral soil, Cudjoe Key, South Florida.

Legit A. H. Curtis.

February.

small yellow flowers (det. L. Brown)

Collected by Pierre Ventur, Yale Department of Anthropology
Improved Performance With Field Element Identifiers

![Graph showing improved performance with FEI encoding]

F-Score Difference vs. Elements
Polygala ambigua, Nutt., var.

Coral soil, Cudjoe Key, South Florida.

Legit A. H. CURTIS.

February.
Learning w/ pre categorization

Gold Labels → Categorization → Class 1 Labels → Machine Learner → Model 1

Class 2 Labels → Machine Learner → Model 2

Class n Labels → Machine Learner → Model n

Unclassified Labels → Categorization → Class 1 Labels → Machine Classification → Classified Labels

Class 2 Labels → Machine Classification → Classified Labels

Class n Labels → Machine Classification → Classified Labels
Specialist100 Curtiss VS 100 General
Future Work

• Community Learning Models
• Label records might be processed in different orders to maximize learning and minimize error rate.
• OCR correction might be improved using context dependent information. Context dependent correction means conducting the correct after knowing the word’s class. For example, word “Ourtiss” should be corrected as “Curtiss”. If the system already identified “Ourtiss” as collector, we can use the smaller collector dictionary instead of using a much larger general dictionary to do the correction.
Community Learning Models

Training Phase

Unclassified Labels

Human Editing

Gold Classified Labels

Machine Learner

Trained Model

Application Phase

Unclassified Labels

Segmentation

Segmented Text

Machine Classifier

Silver Classified Labels
Many thanks to Qin Wei

Listen → Human Learning → Ask Questions