Tri-Trophic Digitization: Putting the OCR in Workflow

Plants, Herbivores, and Parasitoids

NSF ADBC Digitization TCN

Kimberly Watson

iDigBio Augmenting OCR Workshop
October 1, 2012

Project Coordinator at the New York Botanical Garden for the Tri-Trophic Digitization TCN.

Topic: overall workflow for digitizing a large collection of plants and insects across several collaborating institutions, and how we will be using OCR software in the process.

A Tri-Trophic Example

Plants	Herbivores	Parasitoids
Crop Plants	**Hemiptera (e.g. Aphids)**	**Hymenoptera (Parasitoid wasps)**
Produce fruits and tubers of significant agricultural and economic importance.	Pierce plant stems and leaves; specialize on one species or numerous.	Lay eggs inside aphid; larva consumes host from the inside out; emerges from "mummy" as an adult.
Poaceae: corn, wheat, rice **Fabaceae:** soybean, hay **Solanaceae:** tomato, potato	Reduce plant vigor, transmit disease, reduce harvest yield.	

Photo: www.alexanderwild.com

Photo: www.alexanderwild.com

The primary goal of the Tri-Trophic TCN is to digitize, integrate, and make available online the data of three major groups of organisms:

- The Hemiptera, which are a large group of insect herbivores, including mealy bugs and aphids
- The plant taxa commonly eaten by Hemiptera, many of which are important economically and agriculturally
- And the parasitoid Hymenoptera, which parasitize the herbivores by laying their eggs inside them and thereby killing the herbivore.

Species of Interest: North American Biota

Plants

Family	# species
Apiaceae	250
Asteraceae	2,400
Chenopodiaceae	250
Cupressaceae	30
Cyperaceae	850
Fabaceae	850
Fagaceae	97
Grossulariaceae	53
Juglandaceae	17
Lamiaceae	240
Oleaceae	35
Pinaceae	66
Poaceae	1,400
Polygonaceae	440
Rhamnaceae	75
Rosaceae	360
Salicaceae	123
Scrophulariaceae	430
Solanaceae	85
Zygophyllaceae	15
Total	**8,066**

Herbivores

Hemiptera	# species
Coccoidea (scale insects)	986
Aphidoidea (plant lice)	1,532
Psylloidea (jumping plant lice)	176
Auchenorrhyncha (cicadas, hoppers)	4,629
Heteroptera	3,827
Total	**11,150**

Parasitoids

Hymenoptera	# species
Aphelinidae	212
Encyrtidae	490
Mymaridae	187
Signiphoridae	19
Trichogrammatidae	131
Total	**1,039**

Over the course the project, 14 botanical institutions will digitize herbarium specimens from the United States, Canada, and Mexico, representing 20 different plant families and including just over 8000 species, while 18 entomological institutions will digitize insect collections from the same geographic region representing just over 11,000 species of Hemiptera and 1000 species of Hymenoptera.

Insect Specimen Digitization

Institutions (18)	Specimens databased	% Georeferenced	Prior funding	Specimens to be databased
American Museum of Natural History	30,000	100	NSF-PBI	333,000
B. P. Bishop Museum, Honolulu	0	0		70,000
California Academy of Sciences	4,000	100	NSF-PBI	40,000
California Dept. Food & Agriculture	1,000	100	NSF-PBI	75,000
Carnegie Museum, Pittsburgh	0	1		15,000
Colorado State University	0	1		15,000
Cornell University	0	1		30,000
Illinois Natural History Survey	36,000	100	NSF-REVSYS	73,000
Mississippi State University	0	0		50,000
North Carolina State University	1,000	100	NSF-BRC	75,000
Oregon State University	1,000	100		40,000
Texas A&M University	15,000	100	NSF-PBI	150,000
Univ. of California, Berkeley, Essig Museum	12,000	92	NSF-PBI, NSF-BRC	45,000
University of California, Riverside	14,000	100	NSF-PBI, NSF-DBI	75,000
University of Delaware	2,000	0		20,000
University of Kansas	0	0		50,000
University of Kentucky	0	0		35,000
University of Massachussetts, Amherst	10,000	0		15,000
Total	**126,000**			**1,206,000**
Grand Total				**1,332,000**

The entomologists will catalog roughly 1.2 million specimens by manually entering complete collection data into a centralized database hosted by AMNH which allows for remote access and data entry via an online interface. At this time, I don't anticipate that they will be using OCR software for specimen data capture.

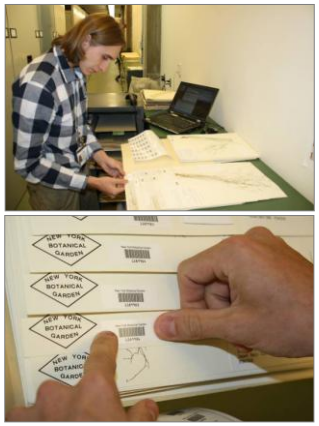
Plant Specimen Digitization

Institutions (14)	Specimens databased	% Georeferenced	Prior funding	Specimens to be databased
Eastern Michigan University	0	0		10,000
Illinois Natural History Survey	308,000	17		94,000
Iowa State University	46,000	0		102,000
Miami University	14,000	5		35,000
Missouri Botanical Garden	247,000	25	NSF-BRC	101,000
New York Botanical Garden	102,000	30	NSF-BRC, NSF-PBI	274,000
University of Colorado	51,000	0		67,000
University of Illinois	0	0		30,000
University of Kansas	129,000	65		97,000
University of Maine	100,000	0		34,000
University of Michigan	26,000	0		115,000
University of Minnesota	93,000	10	NSF- BRC	70,000
University of Texas	105,000	10		105,000
University of Wisconsin	120,000	50		90,000
Total	**1,341,000**			**1,224,000**
GRAND TOTAL				**2,565,000**

The botanists will catalog 1¼ million herbarium specimens, adding to their existing 1¼ million complete specimen records, yielding a total of more than 2½ million records.
Because many herbarium specimen labels are typed or printed, we will not initially keystroke complete collection label data into a database, but instead use OCR software to transcribe data from digital images.

Rapid Data Entry

- **Catalog skeletal records**
  - Barcode
  - Scientific ("Filed As") name
  - Use Tropicos® authority files
- **Average ±150-200/hr**
- **Send existing data to NY**
  - Complete records
  - Georeferenced (if available)
  - Darwin Core format



The workflow begins with each herbarium creating a skeletal database of their specimens, meaning each collection label on a sheet is given a barcode and a skeletal database record containing at least that barcode # and the scientific name under which the specimen is filed in the herbarium.

On average at NY we generate ±150-200 skeletal records per hour.

Additionally, the participating herbaria will send to NY an export of their existing complete specimen records in Darwin Core format.

Rapid Image Capture

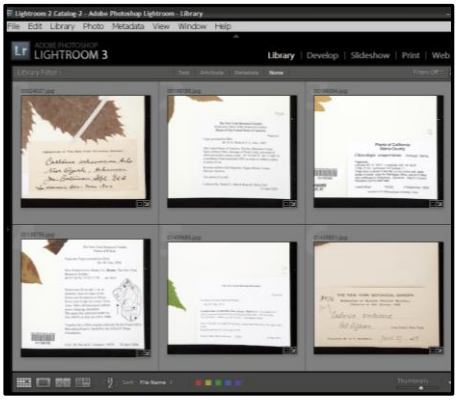
- **Photograph every specimen**
  - 21 megapixel DSLR camera
  - Macro lens, 55 mm
  - Photo-Box, even illumination
- **Barcode = Image file name**
- **Average ±80-120/hr**
- **Send JPG images to NY**

Next, every specimen is photographed using a 21 MP DSLR camera, and the image file renamed as the barcode number. At NY we capture on average ±90-120 exposures per hour. In the end, each herbarium will retain a set of archival images and send a set of JPG derivatives to NY.

Batch Image Post-Processing

JPG images compiled at NY

- **>1 barcode per sheet**
- **Crop to lower right**
- **Crop to label**
- **Export JPGs of labels**



As the JPGs arrive at NY, they are batch imported into Adobe Photoshop Lightroom.

- First, the specimen images having more than 1 barcode (i.e. > 1 collection) are flagged to be handled individually.
- Next, the remaining images are batch-cropped to the lower right corner where the label usually occurs.
- If after cropping the label is not captured, then those images are cropped individually.

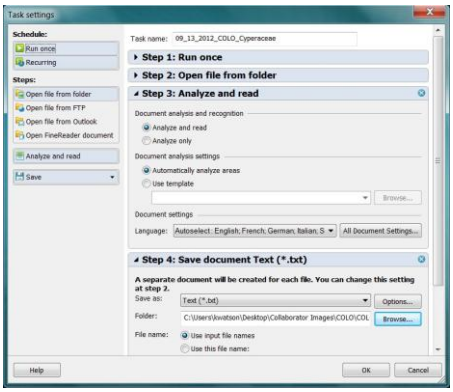
Once a batch is cropped, the label images are Exported as JPGs.

Batch OCR

ABBYY FineReader 11 Corporate Edition

ABBYY Hot Folder

- ✓ Run Once/Recurring
- ✓ Automatically Analyze
- ✓ Autoselect Language
- ✓ Save as text files
  Barcode.txt

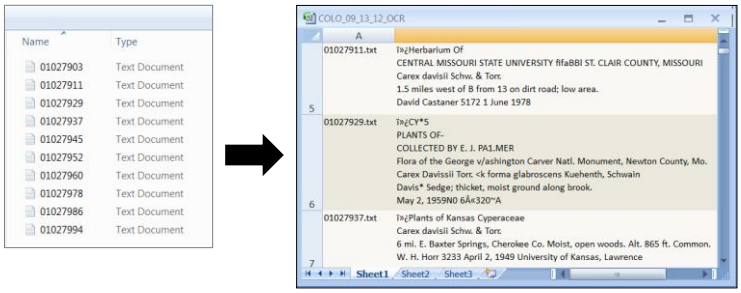


These JPGs are then run through the OCR software. At NY, we opted to use ABBYY FineReader, Corporate Edition, as it allows for batch processing with the Hot Folder.

- Set it to run at a particular time (or have it continually running)
- Direct it to the folder containing the label images
- Select to have the software automatically analyze each image individually and automatically select the language.
- Select to have each file saved as an individual text file, retaining the barcode number as the file name.

Using the OCR data

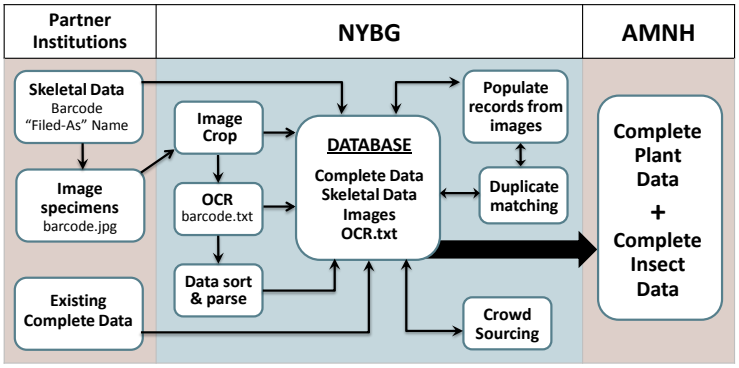
- Merge individual text files into single Excel worksheet using a Powershell script
- Search, group, enter data for several collections at once



Then, using a Powershell script, the individual text files are merged into a single Excel worksheet, where the data can be:

- Searched,
- Grouped (e.g. by Collector, State, etc.)
- And finally, the data parsed into fields for import into the main database.

Plant Specimen Digitization Workflow



- **Complete and skeletal records combined at NYBG**
- **Populate skeletal records using OCR data, duplicate matching, crowd sourcing**

The overall botanical specimen workflow can be diagrammed as shown. Not all the collection labels will yield a lot of OCR data, or perfectly clean data that can be easily parsed; however, at least in many cases it will help to organize the images for more efficient data capture.

Ideally we would like to be able to automate the parsing of OCR data in batches by way of Natural Language Processing (NLP), but this option is not yet available.

In addition to using the OCR output to help organize the images and populate some fields, skeletal records will be populated using

- Duplicate matching with complete records from the combined dataset, or with records from GBIF (=Scatter, Gather, Reconcile)
- And some (i.e. all handwritten labels) will have to be keystroked, we hope with the help of Crowd Sourcing, volunteers, or interns.

Once the plant records are complete, they will be sent to Katja Seltmann, the overall Project Manager at AMNH, who will combine them with the insect data to be georeferenced.

Tri-Trophic TCN Partners

BOTANY

- **Robert Naczi, New York Botanical Garden**
- **Robert Magill, Missouri Botanical Garden**
- **Richard Rabeler, University of Michigan**
- **Melissa Tulig, New York Botanical Garden**
- Barbara Thiers, New York Botanical Garden
- Kim Watson, New York Botanical Garden
- Margaret Koopman, Eastern Michigan University
- Loy Phillippe, Illinois Natural History Survey
- Deborah Lewis, Iowa State University
- Michael Vincent, Miami University
- Timothy Hogan, University of Colorado
- Mary Ann Feist, University of Illinois
- Craig Freeman, University of Kansas
- Christopher Cambell, University of Maine
- Anita Cholewa, University of Minnesota
- Beryl Simpson, University of Texas
- Kenneth Cameron, University of Wisconsin

Data Contributors

- Consortium of Pacific Northwest Herbaria
- Consortium of California Herbaria
- Southwest Biodiversity Consortium

ENTOMOLOGY

- **Randall Schuh, American Museum of Natural History**
- **Christine Johnson, American Museum of Natural History**
- **Christiane Weirauch, University of California, Riverside**
- **John Heraty, University of California, Riverside**
- **Charles Bartlett, University of Delaware**
- **Benjamin Normark, University of Massachusetts, Amherst**
- Katja Seltmann, American Museum of Natural History
- Neal Evenhuis, BP Bishop Museum, Honolulu
- David Kavanaugh ,California Academy of Sciences
- Stephen D. Gaimari ,California Dept. Food and Agriculture
- Chen Young, Carnegie Museum, Pittsburg
- Boris C. Kondratieff, Colorado State University
- James K. Liebherr, Cornell University
- Dmitry Dmitriev, Illinois Natural History Survey
- Richard Brown, Mississippi State University
- Andy Deans, North Carolina State University
- David Maddison, Oregon State University
- Christopher Marshall, Oregon State University
- John Oswald, Texas A&M University
- Kipling Will, University of California, Berkeley
- Caroline Chaboo , University of Kansas
- Michael Sharkey , University of Kentucky
- John Pickering, University of Georgia

Data Contributors

- Canadian National Collection, Ottawa
- University of California, Davis
- Kansas State University

NSF Award#1115104