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.

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.

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

<table>
<thead>
<tr>
<th>Institutions</th>
<th>Specimens databased</th>
<th>% Georeferenced</th>
<th>Prior funding</th>
<th>Specimens to be databased</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Museum of Natural History</td>
<td>30,000</td>
<td>100</td>
<td>NSF-PBI</td>
<td>333,000</td>
</tr>
<tr>
<td>B. P. Bishop Museum, Honolulu</td>
<td>0</td>
<td>0</td>
<td></td>
<td>70,000</td>
</tr>
<tr>
<td>California Academy of Sciences</td>
<td>4,000</td>
<td>100</td>
<td>NSF-PBI</td>
<td>40,000</td>
</tr>
<tr>
<td>California Dept. Food &amp; Agriculture</td>
<td>1,000</td>
<td>100</td>
<td>NSF-PBI</td>
<td>75,000</td>
</tr>
<tr>
<td>Carnegie Museum, Pittsburgh</td>
<td>0</td>
<td>1</td>
<td></td>
<td>15,000</td>
</tr>
<tr>
<td>Colorado State University</td>
<td>0</td>
<td>1</td>
<td></td>
<td>15,000</td>
</tr>
<tr>
<td>Cornell University</td>
<td>0</td>
<td>1</td>
<td></td>
<td>30,000</td>
</tr>
<tr>
<td>Illinois Natural History Survey</td>
<td>36,000</td>
<td>100</td>
<td>NSF-REV/SYS</td>
<td>73,000</td>
</tr>
<tr>
<td>Missouri State University</td>
<td>0</td>
<td>0</td>
<td></td>
<td>50,000</td>
</tr>
<tr>
<td>North Carolina State University</td>
<td>1,000</td>
<td>100</td>
<td>NSF-PBI</td>
<td>75,000</td>
</tr>
<tr>
<td>Oregon State University</td>
<td>1,000</td>
<td>100</td>
<td></td>
<td>40,000</td>
</tr>
<tr>
<td>Texas A&amp;M University</td>
<td>15,000</td>
<td>100</td>
<td>NSF-PBI</td>
<td>150,000</td>
</tr>
<tr>
<td>Univ. of California, Berkeley, Eng. Museum</td>
<td>12,000</td>
<td>92</td>
<td>NSF-PBI, NSF-BRC</td>
<td>45,000</td>
</tr>
<tr>
<td>University of California, Riverside</td>
<td>14,000</td>
<td>100</td>
<td>NSF-PBI, NSF-EBI</td>
<td>75,000</td>
</tr>
<tr>
<td>University of Delaware</td>
<td>2,000</td>
<td>0</td>
<td></td>
<td>20,000</td>
</tr>
<tr>
<td>University of Kansas</td>
<td>0</td>
<td>0</td>
<td></td>
<td>50,000</td>
</tr>
<tr>
<td>University of Kentucky</td>
<td>0</td>
<td>0</td>
<td></td>
<td>35,000</td>
</tr>
<tr>
<td>University of Massachusetts, Amherst</td>
<td>10,000</td>
<td>0</td>
<td></td>
<td>15,000</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>126,000</strong></td>
<td></td>
<td></td>
<td><strong>1,206,000</strong></td>
</tr>
<tr>
<td><strong>Grand Total</strong></td>
<td><strong>1,332,000</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Institutions</th>
<th>Specimens databased</th>
<th>% Georeferenced</th>
<th>Prior funding</th>
<th>Specimens to be databased</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eastern Michigan University</td>
<td>0</td>
<td>0</td>
<td></td>
<td>10,000</td>
</tr>
<tr>
<td>Illinois Natural History Survey</td>
<td>308,000</td>
<td>17</td>
<td></td>
<td>94,000</td>
</tr>
<tr>
<td>Iowa State University</td>
<td>46,000</td>
<td>0</td>
<td></td>
<td>102,000</td>
</tr>
<tr>
<td>Missouri University</td>
<td>14,000</td>
<td>5</td>
<td></td>
<td>35,000</td>
</tr>
<tr>
<td>Missouri Botanical Garden</td>
<td>247,000</td>
<td>25</td>
<td>NSF-BRC</td>
<td>101,000</td>
</tr>
<tr>
<td>New York Botanical Garden</td>
<td>102,000</td>
<td>30</td>
<td>NSF-BRC, NSF-PBI</td>
<td>274,000</td>
</tr>
<tr>
<td>University of Colorado</td>
<td>51,000</td>
<td>0</td>
<td></td>
<td>87,000</td>
</tr>
<tr>
<td>University of Illinois</td>
<td>0</td>
<td>0</td>
<td></td>
<td>30,000</td>
</tr>
<tr>
<td>University of Kansas</td>
<td>129,000</td>
<td>65</td>
<td></td>
<td>97,000</td>
</tr>
<tr>
<td>University of Maine</td>
<td>100,000</td>
<td>0</td>
<td></td>
<td>34,000</td>
</tr>
<tr>
<td>University of Michigan</td>
<td>36,000</td>
<td>0</td>
<td></td>
<td>115,000</td>
</tr>
<tr>
<td>University of Minnesota</td>
<td>93,000</td>
<td>10</td>
<td>NSF-BRC</td>
<td>70,000</td>
</tr>
<tr>
<td>University of Texas</td>
<td>105,000</td>
<td>10</td>
<td></td>
<td>105,000</td>
</tr>
<tr>
<td>University of Wisconsin</td>
<td>130,000</td>
<td>50</td>
<td></td>
<td>90,000</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1,341,000</strong></td>
<td></td>
<td></td>
<td><strong>1,224,000</strong></td>
</tr>
<tr>
<td><strong>GRAND TOTAL</strong></td>
<td><strong>2,565,000</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.
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.

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.

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

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.

Plant Specimen Digitization Workflow

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

Tri-Trophic TCN Partners

- Robert Naczi, New York Botanical Garden
- Robert Magill, Missouri Botanical Garden
- Richard Betzler, University of Michigan
- Melissa Tully, New York Botanical Garden
- Barbara Thorns, New York Botanical Garden
- Katherine Koppman, Eastern Michigan University
- Jay Phillips, Illinois Natural History Survey
- Deborah Lewis, Iowa State University
- Mitchell Vincent, Miami University
- Timothy Hage, University of Colorado
- Mary Anne Field, University of Illinois
- Craig Freeman, University of Kansas
- Christopher Campbell, University of Maine
- Anita Chaboo, University of Minnesota
- Beryl Lequeux, University of Texas
- Kristoffe Cameron, University of Wisconsin

Data Contributors:
- Consortium of Pacific Northwest Herbaria
- Consortium of California Herbaria
- Southwest Botany Consortium

ENTOMOLOGY
- Randall Schuh, American Museum of Natural History
- Christine Johnson, American Museum of Natural History
- Christiane Weitschies, University of California, Riverside
- John Horty, University of California, Riverside
- Charles Bartlett, University of Delaware
- Benjamin Norman, University of Massachusetts, Amherst
- Katja Seltmann, American Museum of Natural History
- Neal Evermann, JP Bishop Museum, Honolulu
- David Kevans, California Academy of Sciences
- Stephen D. Ginger, California Dept. Food and Agriculture
- Chen Young, Carnegie Museum, Pittsburgh
- Boris C. Konstantin, Colorado State University
- James E. Laidet, Cornell University
- Dmitry Ovchinnikov, Illinois Natural History Survey
- Richard Brown, Mississippi State University
- Andy Dean, North Carolina State University
- David Mallison, Oregon State University
- Christopher Marshall, Oregon State University
- John Oswald, Texas A&M University
- Epling Will, University of California, Berkeley
- Caroline Chace, University of Kansas
- Michael Sharkey, University of Kentucky
- John Pohler, University of Georgia

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

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.