1

Tri-Trophic Digitization Strategies

Plants, Herbivores, and Parasitoids

NSF ADBC Digitization TCN

Kimberly Watson, Melissa Tulig

iDigBio Botany 2012 Digitization Workshop
July 12, 2012

Kim Watson (presenting): Project Coordinator at the New York Botanical Garden for the Tri-Trophic Digitization TCN.

Melissa Tulig: Principal Investigator at the New York Botanical Garden for the Tri-Trophic Digitization TCN.

Subject: Strategies in digitizing a large collection of plants and insects across several collaborating institutions.

2

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

In brief, the goal of the Tri-Trophic Digitization TCN is to digitize and integrate the data of three major groups of organisms:

- The Hemiptera, a large group of herbivores including many common garden pests, such as mealy bugs, aphids, and armored scales.
- The plant taxa which are commonly eaten by the Hemiptera, many of which are important economically and agriculturally.
- And the parasitoid Hymenoptera, which parasitize the Hemiptera by laying their eggs inside them, which then develop into larvae that consume the herbivore from the inside out.

3

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 of 4 years, we intend to digitize a large number of species.

The collaborating botanical institutions will digitize herbarium specimens from 20 plant families, amounting to just over 8000 species.

And the entomological institutions will digitize insect collections from just over 11,000 species of Hemiptera and 1000 species of Hymenoptera.

Given the obvious differences between these types of collections, we are implementing different digitization workflows depending on the trophic level, streamlined wherever possible.

4

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**

There are 14 collaborating botanical institutions, contributing over 1¼ million complete specimen records.

And over the course of the project, we aim to digitize another 1¼ million specimens, yielding a total of more than 2½ million herbarium specimen records.

5

Plant Specimen Digitization

Streamlined Workflow for Rapid Data Entry

Participating herbaria will

- **Catalog skeletal records for all collections**
  - Barcode
  - Scientific ("Filed As") name
  - Use Tropicos® authority files from the Missouri Botanical Garden
- **Send existing data export to NY**
  - Complete records
  - Georeferenced (if available)
  - Darwin Core format

For all participating herbaria, the first steps in the workflow are to curate their collections and then catalog skeletal records for each collection.

Every collection label is given a barcode, applied in rapid fashion, and a skeletal database record containing the barcode number and the scientific name under which the specimen is filed in the herbarium.

To standardize these scientific names across all institutions, we are using an export of TROPICOS authority files provided by the Missouri Botanical Garden.

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

Each participating herbarium will also be sending to NYBG an export of their existing complete specimen records in Darwin Core format.

6

Plant Specimen Digitization

Streamlined Workflow for Rapid Image Capture

Participating herbaria will

- **Purchase imaging equipment**
  - 21 megapixel DSLR camera
  - Macro autofocus lens, 55 mm
  - Photo-eBox, even illumination
- **Schedule Site-Visit**
  - Assembly and training
- **Photograph every specimen**
  - Barcode = Image file name
- **Send JPG images to NY**

The next step is for each institution to photograph every barcoded specimen.

First, each institution will order imaging equipment that includes a 21 MP DSLR camera, macro lens, and Photo-ebox.

Once all their equipment has arrived, the NYBG Project Coordinator will make a 2-day site-visit to assemble the equipment and train all the participants in how to photograph their specimens and manage the thousands of image files they'll be generating.

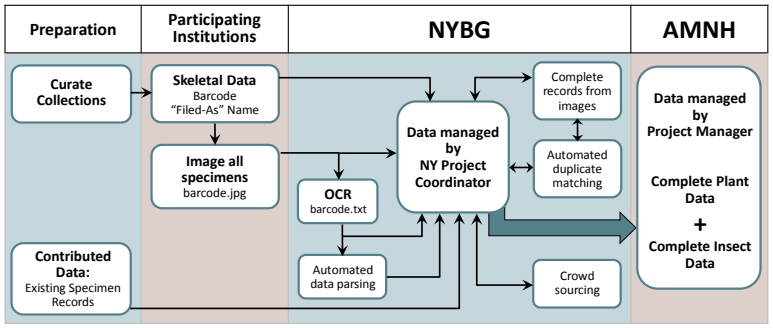
At NY we capture an average ±90-120 exposures per hour.

Each image is named with the barcode number.

In the end, each herbarium will have a set of archival images, and they will send a set of JPG copies to NYBG.

7

Plant Specimen Digitization



- All plant data and images will be compiled at NYBG
- Run OCR software on collection labels
- Skeletal records: Barcode, "Filed-As" name, JPG image, editable OCR text
- NY Project Coordinator will complete all partial records

As NYBG receives all this data, including complete records, skeletal records, and JPG image files from the other institutions, the NYBG Project Coordinator will use whatever tools are available to complete all the partial records.

All collection labels will be OCR'd; the hope is that many will have data that can be auto-parsed.

Some records will be completed using duplicate matching.

And some records will have to be key-stroked, either by the NYBG Project Coordinator, or through the help of crowd sourcing, interns, volunteers, and citizen scientists.

Once the plant records are complete, they are sent to Katja Seltmann, the Tri-Trophic Digitization Project Manager at AMNH.

8

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**

There are 18 participating entomological institutions, contributing over 100,000 complete specimen records, many of which are georeferenced.

Over the course of the project, they aim to digitize another 1.2 million specimens, yielding a total of over 1.3 million insect specimen records.

9

Insect Specimen Digitization

Streamlined Workflow for Rapid Data Entry

Curate and stage specimens
- Scientific name (determined by specialist)
- Collection event
- Sex

Pin barcode to each specimen

Enter complete specimen label and host data

Similar to the workflow discussed before, the first step in the insect workflow involves the curation and staging of the specimens.

At AMNH, the specimens are staged one unit-tray at a time, organized first by scientific name (as determined by a specialist), then by collection event, followed by sex.

Once organized, a tiny barcode is pinned to each collection, and then complete specimen label data is key-stroked into the database, including sex and host information.

10

Insect Specimen Digitization

Streamlined Web Interface for Rapid Data Entry

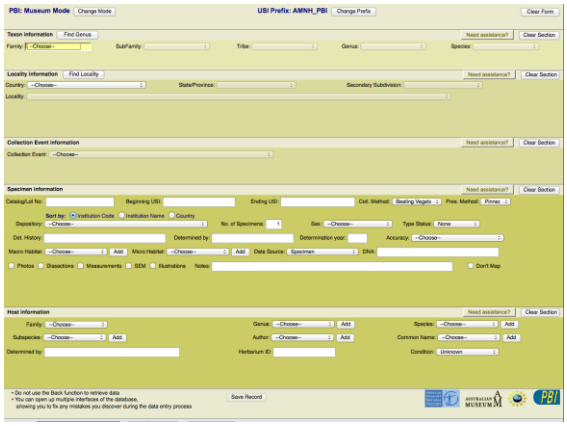
Taxon

Locality

Collection event

Specimen data

Host data
- Plant
- Herbivore



The database into which most of the insect records will be entered is one hosted by AMNH which allows for remote access and data entry via an online interface.

This single-page interface with look-up-lists allows for rapid data entry with login access from anywhere with internet access, and it is easy to use by new, untrained personnel.

The authority files for taxon and host data have all been compiled and standardized from various sources by Project Manager, Katja Seltmann, so now the database includes all those for Hemiptera, Hymenoptera, and any host plant taxa.

11

Insect Specimen Digitization

Imaging

- Image representative specimens for each species
- Use existing imaging stations at partner institutions
- About 30% of Hemiptera are complete
- Expect to produce about 20,000 new images

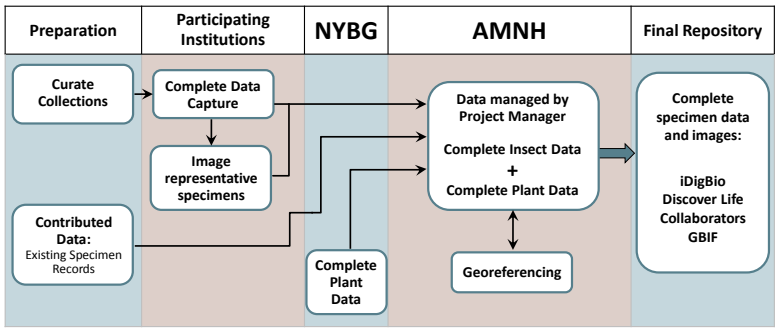
Unlike for the botanical collections, only representative insect specimens for each species will be imaged for the project.

These images will be captured using existing imaging stations at collaborating institutions.

We expect to produce approximately 20,000 new high-resolution, close-up images.

12

Insect Specimen Digitization



- All data aggregated at AMNH
- Project Manager, Katja Seltmann (AMNH), will manage data and georeferencing
- Completed plant and insect data submitted to iDigBio, Discover Life, GBIF, etc.
- Completed plant data returned to participating herbaria

Although the initial rate of specimen data capture will be slower for the insects, they will be generating complete records from the start.

The Project Manager, Katja Seltman, will manage all the insect data that is accumulated for the project, a task made easier by the centralized insect database hosted at AMNH.

She will then combine the insect data set with the completed plant data sent from NYBG.

Once aggregated, the combined data sets will be georeferenced. At the end of the project, the combined data set will be submitted to Discover Life, iDigBio, etc.

And the completed botanical records will be returned to the participating herbaria.

13

Digitization Challenges

OVERALL:
- Insure accuracy of specimen identifications
- Implement authority files for all groups in all databases
- Integrate data across databases
- Maintain data over the long-term

PLANTS:
- Duplicates with differing names; how to report discrepancies to collaborators
- Train collaborators to manage data and images, use imaging equipment
  - Digitization and data management experience
  - Technical support
- Long-term archival image storage for all institutions? 36+ TB of raw files

INSECTS:
- Expand existing database to include authority files for parasitoids, and plants
- How to transfer data from images of scanned microscope slides

As streamlined and well vetted as these workflows are, there remain some significant challenges. Overall, some such challenges include:

- How to insure the accuracy of specimen identifications.
- How to implement authority files for all groups in all databases
- How to integrate data across databases
- How to maintain this data over the long-term

PLANTS: Duplicate specimens with differing determinations; training collaborators with varying degrees of data management experience and technical support.

INSECTS: Had to expand the existing database to include parasitoid and plant authority files; how to digitize insect collection on microscope slides.

14

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

Many thanks to all the collaborating institutions, data contributors, iDigBio for hosting the Botany 2012 Digitization Workshop, and, of course, to NSF for the funding and opportunity to participate in such an incredible project.