

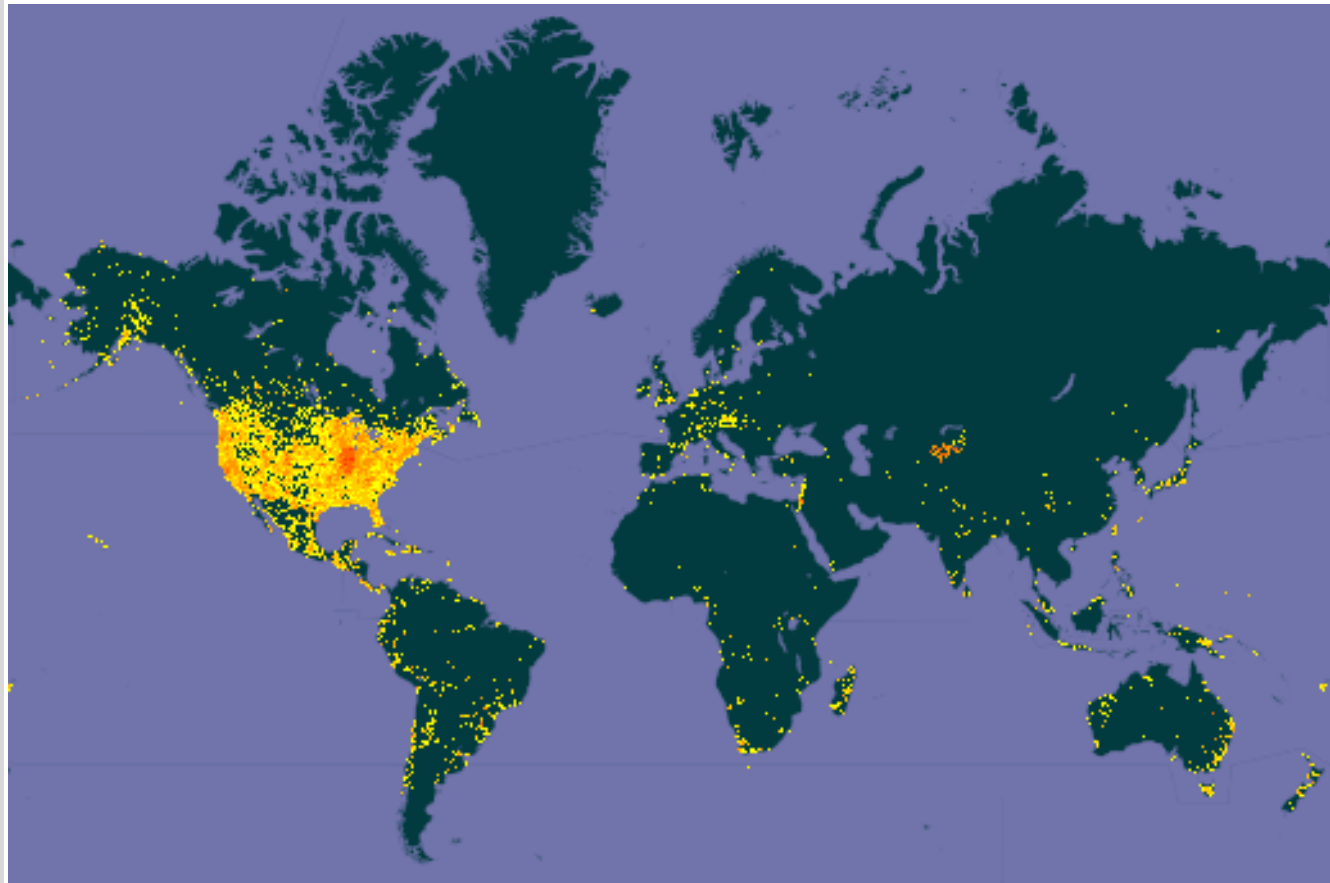
# Getting an insect collection ready for digitization at the Illinois Natural History Survey's InvertNet project

Dmitry Dmitriev



# Illinois Natural History Survey collection

- > 7,000,000 specimens
- > 14,000 primary types



Today, ~900,000 records (2.1 million specimens)

GBiF: <http://data.gbif.org/datasets/provider/75>

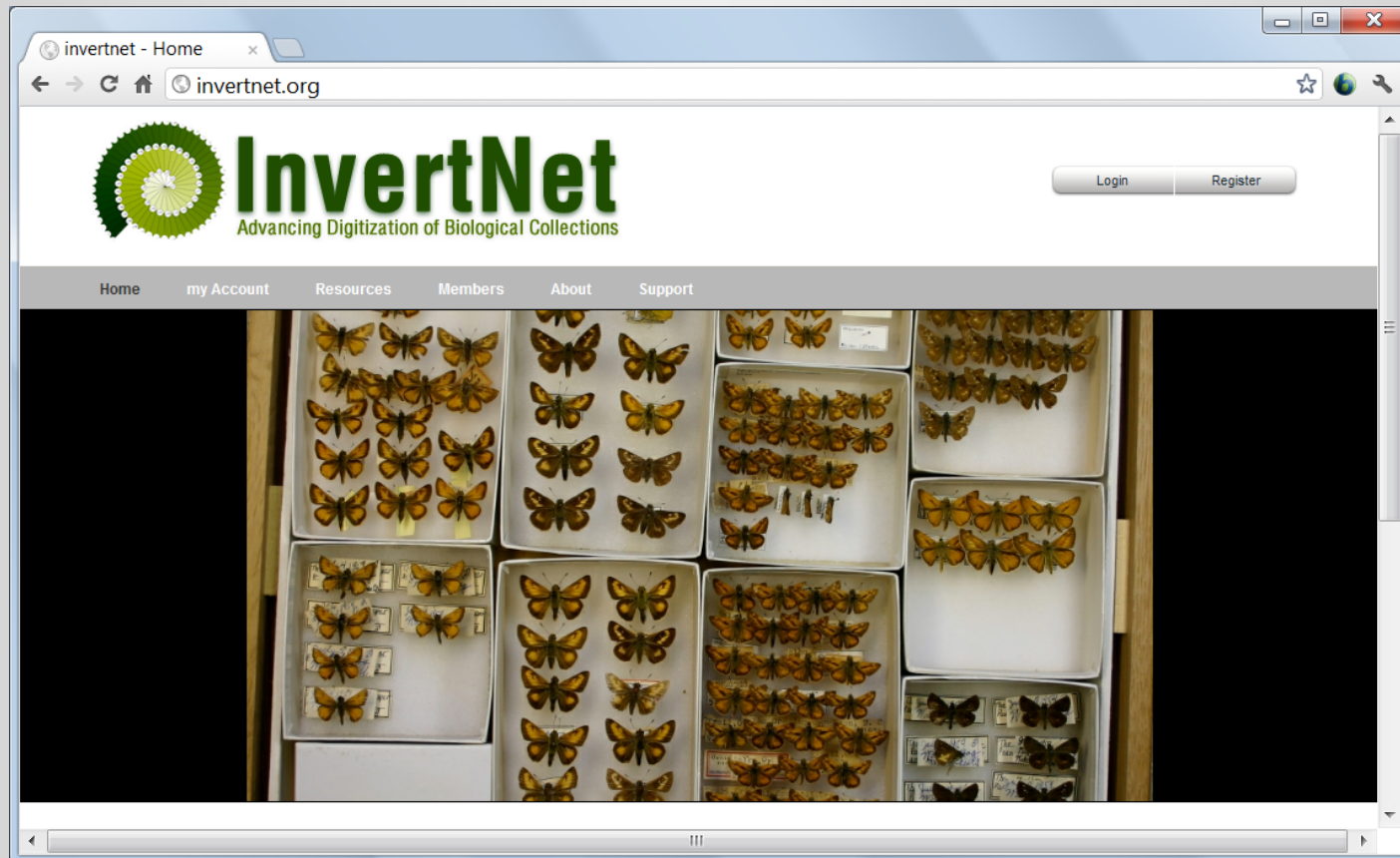
# InvertNet Rationale

- vast majority of specimens the collections are invertebrates
  - primarily insects and related arthropods
  - less than 5% available online
  - only label data usually provided
- existing digitization methods are inadequate
  - slow and expensive (\$1+ per specimen)
  - risk of damage to specimens from handling

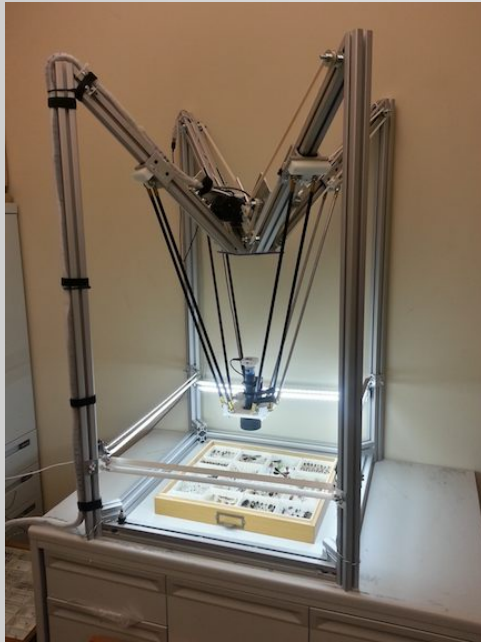


# InvertNet

- Digitize all holdings of 22 arthropod collections (>50 million specimens)
- Provide access to images and other data via online virtual museum
- Provide platform for research and development of additional tools and resources



# Drawer Imaging



- Delta Robot, digital camera, telecentric lens captures grid of single, close-up images at 40-60 x/y coordinates and 5 perspectives



- Single images stitched to yield Gigapixel images from multiple viewpoints



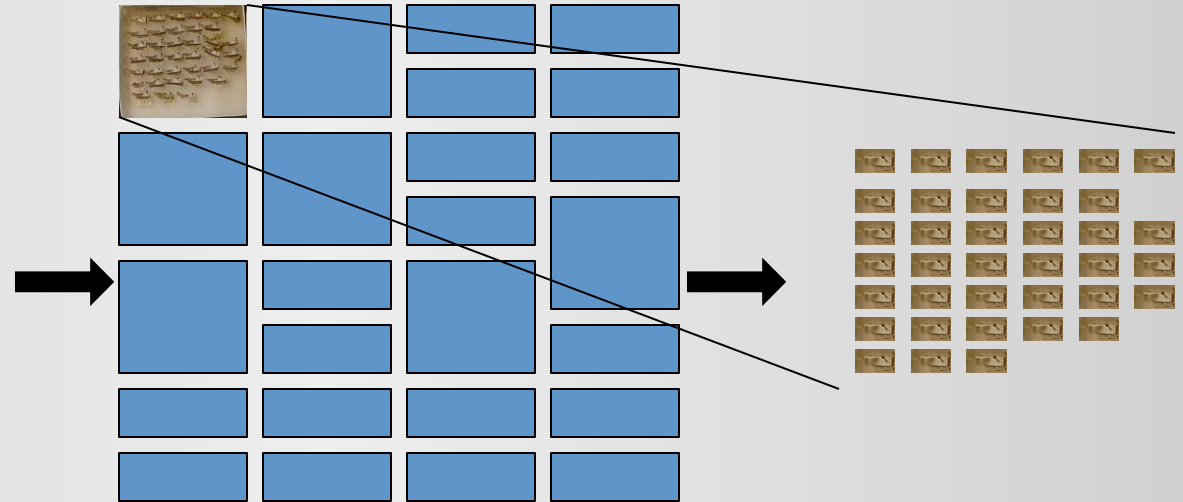
Top-down view



Angled view

- Enables virtual tilting

# Image segmentation/annotation



1. capture image of drawer + metadata (location, contents)

2. segment unit trays (image analysis software)

3. segment specimens  
4. capture label data

# Preparation of the collection for digitization

- Loan returns
- Merging large donations
- Updating nomenclature
- Unifying labels
- Collection profiling

# Loan returns

- In 2011: 876 loans in the collection database, 223 of those cancelled.
- Frustrations: missing records, inaccurate records, human factor.
- Now: 3068 records in the database, 2565 cancelled, 82 renewed. All the invoices scanned.

**ILLINOIS NATURAL HISTORY SURVEY**

607 East Peabody Drive, Champaign, Illinois, USA 61820

Cancelled \_\_\_\_\_

Invoice No. 1827  
Date 16 July 1999

---

<p><b>Upon receipt of material please sign and return one copy of the invoice to K. R. Zeiders</b></p>	Felipe N. Soto Department of Biology University of Vermont  Burlington, VT 05405 f-soto@life.uiuc.edu
--	--

The following specimens have been forwarded as:  (1) loan (2) gift (3) return (4) exchange

Correspondence: per request in person 16 July 1999 Handled by: K. R. Zeiders

---

**HIGHER TAXA** COLLEMBOLA

List of specimens	Access.#	# of vials, etc.	# of spec.	date ret.	Type/country of origin/other
Collembola		55 vials			Kyrgystan



# Large donations

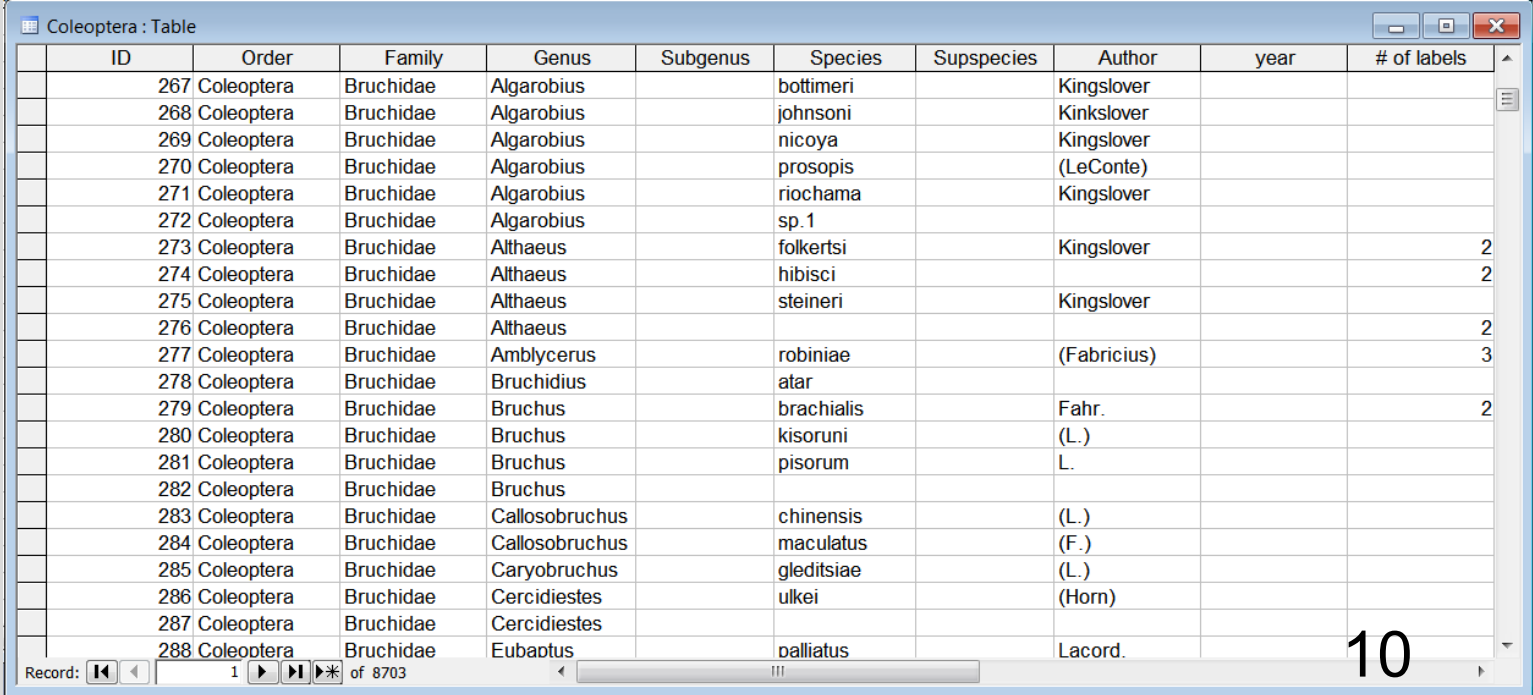
- The material was sitting in temporary storage area, and has not been incorporated in main collection.



# Updating nomenclature

CoL match: Diptera – 60%, Hemiptera – 60%, Lepidoptera – 80%, Coleoptera – 35%.

Now: 63,545 names in the database.



ID	Order	Family	Genus	Subgenus	Species	Supspecies	Author	year	# of labels
267	Coleoptera	Bruchidae	Algarobius		bottimeri		Kingslover		
268	Coleoptera	Bruchidae	Algarobius		johnsoni		Kingslover		
269	Coleoptera	Bruchidae	Algarobius		nicoya		Kingslover		
270	Coleoptera	Bruchidae	Algarobius		prosopis		(LeConte)		
271	Coleoptera	Bruchidae	Algarobius		riochama		Kingslover		
272	Coleoptera	Bruchidae	Algarobius		sp.1				
273	Coleoptera	Bruchidae	Althaeus		folkertsi		Kingslover		2
274	Coleoptera	Bruchidae	Althaeus		hibisci				2
275	Coleoptera	Bruchidae	Althaeus		steineri		Kingslover		
276	Coleoptera	Bruchidae	Althaeus						2
277	Coleoptera	Bruchidae	Amblycerus		robiniae		(Fabricius)		3
278	Coleoptera	Bruchidae	Bruchidius		atar				
279	Coleoptera	Bruchidae	Bruchus		brachialis		Fahr.		2
280	Coleoptera	Bruchidae	Bruchus		kisoruni		(L.)		
281	Coleoptera	Bruchidae	Bruchus		pisorum		L.		
282	Coleoptera	Bruchidae	Bruchus						
283	Coleoptera	Bruchidae	Callosobruchus		chinensis		(L.)		
284	Coleoptera	Bruchidae	Callosobruchus		maculatus		(F.)		
285	Coleoptera	Bruchidae	Caryobruchus		gleditsiae		(L.)		
286	Coleoptera	Bruchidae	Cercidiestes		ulkei		(Horn)		
287	Coleoptera	Bruchidae	Cercidiestes						
288	Coleoptera	Bruchidae	Eubaptus		balliatus		Lacord.		

Record: 1 of 8703

# Unifying labels



*Cistalia signoretii* (Guerin-Meneville 49,288  
1857)

Heteroptera: Lygaeoidea: Rhyparochromidae

# Collection profiling

INHS collection profiling  
(Colin Favret et al. 2007)



Amy Bader  
processing loan return

Form1 : Form

ID:  Person: D. Dmitriev

Room:  Taxon Code:

Label 1

Line 1: THYSANOPTERA

Line 2: Terebrantia

Line 3: Thripidae

Line 4:

Label 2

Line 1: Frankliniella

Line 2: nervosa, musaeperda, minuta

Line 3:

Line 4:

Notes:

Collection Type:

Conservation Status:

Processing State:

Container Condition:

Condition of Labels:

Identification Level:

Arrangement Level:

Data Quality:

Computerization Level:

Number of Specimens:      Print label 1:  Print label 2:

Number of Vials / Slides:

Created: A. Klipp  8/26/2011 1:45:43 PM Modified: D. Dmitriev  2/8/2012 7:45:29 PM

Record:       of 6392

# Collection profiling: Conservation status

Form1 : Form

ID:  Person:

Room:  Taxon Code:   Please select a person from the list

Label 1

Line 1:

Line 2:

Line 3:

Line 4:

Label 2

Line 1:

Line 2:

Line 3:

Line 4:

Notes:

Collection Type:

Conservation Status:

Processing State:

Container Condition:

Condition of Labels:

Identification Level:

Arrangement Level:

Data Quality:

Computerization Level:

Number of Specimens:     Print label 1:  Print label 2:

Number of Vials / Slides:

Created:  Modified:

Record:      of 16449

# Collection profiling: Processing state

Form1 : Form

ID:  Person:

Room:  Taxon Code:   Please select a person from the list

Label 1

Line 1:

Line 2:

Line 3:

Line 4:

Label 2

Line 1:

Line 2:

Line 3:

Line 4:

Notes:

Collection Type:

Conservation Status:

Processing State:

Container Condition: 1 - bulk, unprocessed specimens

Condition of Labels: 2 - specimens lacking labels or not properly prepared

Identification Level: 3 - properly sorted and labeled

Arrangement Level:

Data Quality:

Computerization Level:

Number of Specimens:     Print label 1:  Print label 2:

Number of Vials / Slides:

Created:  Modified:

Record:      of 16449

# Collection profiling: Container condition

Form1 : Form

ID:  Person:

Room:  Taxon Code:   Please select a person from the list

**Label 1**

Line 1:

Line 2:

Line 3:

Line 4:

**Label 2**

Line 1:

Line 2:

Line 3:

Line 4:

Notes:

Collection Type:

Conservation Status:

Processing State:

Container Condition:

Condition of Labels: 1 - cigar boxes, pill boxes, paper bags

Identification Level: 2 - substandard containers, hart botom unit trays

Arrangement Level: 3 - wrong size or dirty unit trays

Data Quality: 4 - archival unit trays

Computerization Level:

Number of Specimens:     Print label 1:  Print label 2:

Number of Vials / Slides:

Created:  Modified:

Record:      of 16449

# Collection profiling: Condition of labels

Form1 : Form

ID: (Number) Person:

Room:  Taxon Code:   Please select a person from the list

**Label 1**

Line 1:

Line 2:

Line 3:

Line 4:

**Label 2**

Line 1:

Line 2:

Line 3:

Line 4:

Notes:

Collection Type:

Conservation Status:

Processing State:

Container Condition:

Condition of Labels:

Identification Level: 1 - faded to illegible, crumbling or missing

Arrangement Level: 2 - partially faded, on non-archival paper

Data Quality: 3 - labels on archival paper

Computerization Level:

Number of Specimens:     Print label 1:  Print label 2:

Number of Vials / Slides:

Created:  Modified:

Record:   16449   of 16449



# Collection profiling: Identification level

Form1 : Form

ID: (Number) Room: Taxon Code: Person:    Please select a person from the list

**Label 1**

Line 1:

Line 2:

Line 3:

Line 4:

**Label 2**

Line 1:

Line 2:

Line 3:

Line 4:

Notes:

Collection Type:

Conservation Status:

Processing State:

Container Condition:

Condition of Labels:

Identification Level:

Arrangement Level:

Data Quality:

Computerization Level:

Number of Specimens:      Print label 1:  Print label 2:

Number of Vials / Slides:

Created:  Modified:

Record:      of 16449

# Collection profiling: Arrangement level

Form1 : Form

ID: (Number) Person:

Room:  Taxon Code:   Please select a person from the list

**Label 1**

Line 1:

Line 2:

Line 3:

Line 4:

**Label 2**

Line 1:

Line 2:

Line 3:

Line 4:

Notes:

Collection Type:

Conservation Status:

Processing State:

Container Condition:

Condition of Labels:

Identification Level:

Arrangement Level:

Data Quality: 1 - mixed taxa in same container  
2 - specimens crowded arranged at higher taxonomic level, or species sharing trays  
3 - specimens arranged alphabetically or phylogenetically  
4 - specimens arranged geographically or numerically within a taxon

Computerization Level:

Number of Specimens:

Number of Vials / Slides:

Created:  Modified:

Record:   16449   of 16449

# Collection profiling: Data quality

Form1 : Form

ID: (Number) Person:

Room:  Taxon Code:   Please select a person from the list

**Label 1** **Label 2**

Line 1:  Line 1:

Line 2:  Line 2:

Line 3:  Line 3:

Line 4:  Line 4:

Notes:

Collection Type:

Conservation Status:

Processing State:

Container Condition:

Condition of Labels:

Identification Level:

Arrangement Level:

Data Quality:

Computerization Level: 1 - data absent often codes only

Number of Specimens: 2 - missing data can be inferred

Number of Vials / Slides: 3 - all data fields intact

4 - value-added data, including retrospective georeferencing

Created:  10/6/2014 7:53:16 AM Modified:  10/6/2014 7:53:16 AM

Record:   16449   of 16449

# Collection profiling: Computerization level

Form1 : Form

ID: (Number)      Person:

Room:     Taxon Code:         **Please select a person from the list**

**Label 1**

Line 1:

Line 2:

Line 3:

Line 4:

**Label 2**

Line 1:

Line 2:

Line 3:

Line 4:

Notes:

Collection Type:

Conservation Status:

Processing State:

Container Condition:

Condition of Labels:

Identification Level:

Arrangement Level:

Data Quality:

Computerization Level:

Number of Specimens: 2 - no computerization

Number of Vials / Slides: 3 - taxonomic information computerized

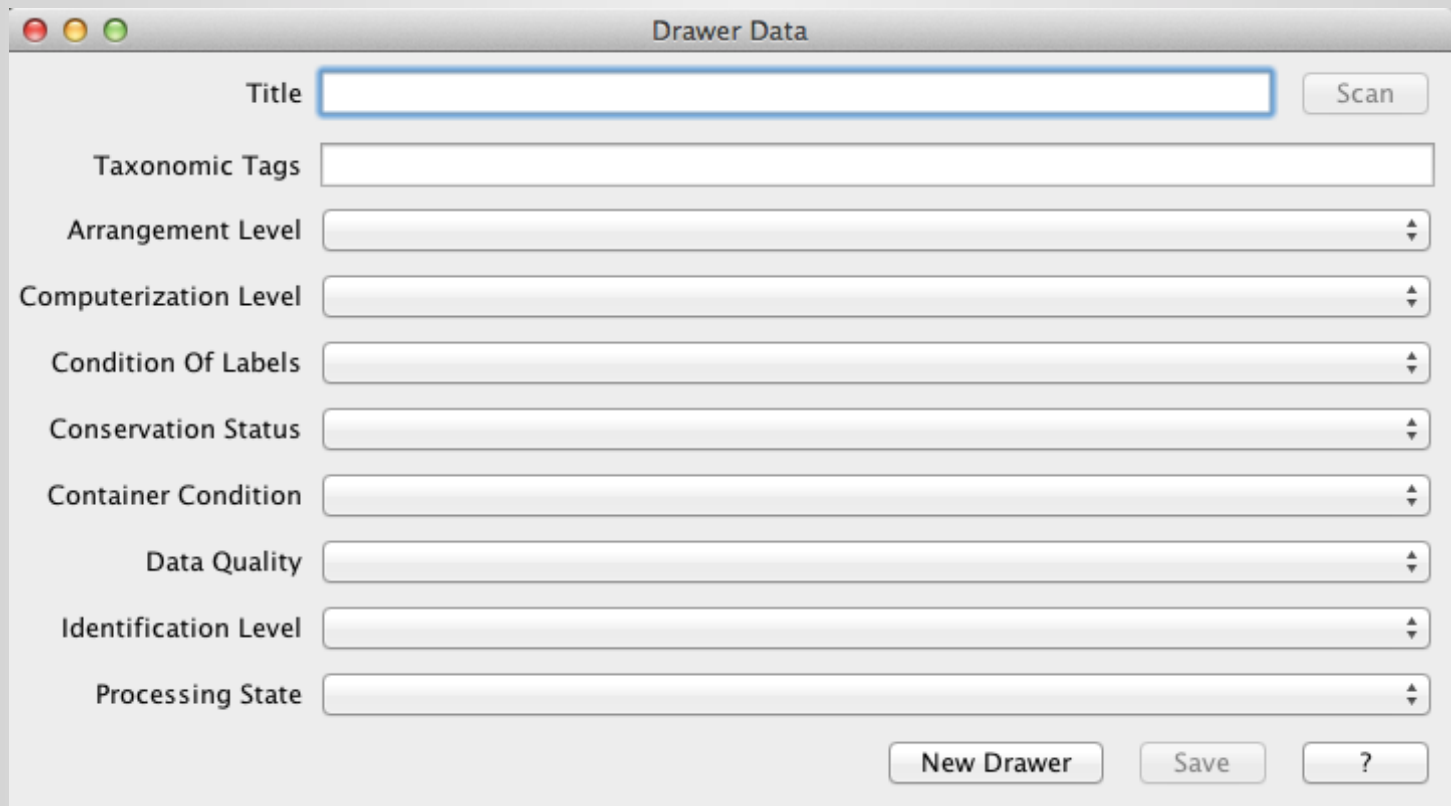
4 - all specimens databased and localities georeferenced

Created:  10/6/2014 7:53:16 AM    Modified:  10/6/2014 7:53:16 AM

Record:   16449   of 16449

# Scanning - Data Entry

- Title must be unique
- After you have entered a title, press Scan
- Tags should include any relevant taxa
- Fill in as many quality menus as you can
- Don't forget to save!
- Once scanning is done, press New Drawer



The screenshot shows a window titled "Drawer Data" with a standard macOS-style title bar (red, yellow, green buttons). The form contains the following fields and controls:

- Title:** A text input field with a blue border, followed by a "Scan" button.
- Taxonomic Tags:** A wide text input field.
- Arrangement Level:** A dropdown menu.
- Computerization Level:** A dropdown menu.
- Condition Of Labels:** A dropdown menu.
- Conservation Status:** A dropdown menu.
- Container Condition:** A dropdown menu.
- Data Quality:** A dropdown menu.
- Identification Level:** A dropdown menu.
- Processing State:** A dropdown menu.

At the bottom of the window, there are three buttons: "New Drawer", "Save", and a help button with a question mark.

# Collection profiling



# Collection profiling



# Label data capture

- OCR doesn't work
  - labels partly obscured by specimens
  - many hand written
  - high error rates—time fixing errors > time required to manually enter data
- Manual entry of verbatim label more accurate
  - still expensive, time consuming, error prone
- Crowd-sourcing is most viable option
  - rapid
  - low cost
  - built-in redundancy to reduce errors
  - applications already available (e.g., Notes from Nature)

Notes from Nature .

COLLECTIONS ABOUT DISCUSS BLOG

Transcribe museum records to

# TAKE NOTES FROM NATURE

START TRANSCRIBING

2 Collections available	10059 Total specimens	0% Transcription progress	3 Users contributing
----------------------------	--------------------------	------------------------------	-------------------------



# Online data entry form

Illinois Natural History x

membracoidea.speciesfile.org/InsectCollectionEdit.aspx

Welcome Dmitry Dmitriev [Sign Out](#)

**Illinois Natural History Survey: Insect Collection**

**Edit Form**

[User Settings](#) | [Delete Specimen Data](#) | [Enter Specimen Data](#)

**Specimen Data Entry**

Find Specimen     
Recent entries.

Catalog Number   -

Preparation Type

Number of Specimens  Males  Females  Adults Unsexed   
 Immatures  Pupae  Exuvia   
 Age Unknown  Others

Taxon Code

Locality Label

Accession Number

Determination Label


Other Label

Comments

Next record  Catalog number +1  Same taxon  
 Same locality  Same other labels

[Clear](#) [New record](#)

ILLINOIS NATURAL HISTORY SURVEY  
PRAIRIE RESEARCH INSTITUTE

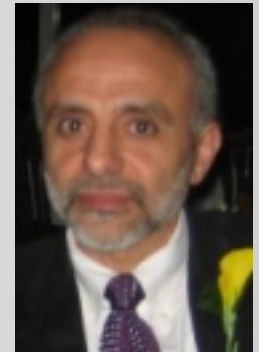


# Parsing verbatim labels

- Matching with existing parsed records (~40% match)
- Finding date (~95% records)
- Finding coordinates (~85% records)
- Finding collector (~70% records)
- Finding identifier (~30% records)
- Student parsing
- Crowd-sourcing (have not been tested yet).

# InvertNet UIUC Team

- Chris Dietrich – Director
  - Systematic Entomologist
- John Hart – CoPI
  - Computer Science - Graphics
- Nahil Sobh – CoPI
  - Computational Multiscale Nanosystems
- Umberto Ravaioli – CoPI
  - Computational Multiscale Nanosystems
- David Raila – Senior Collaborator
  - Computer Science – Sr. Research Programmer
- Others
  - Programmers, research assistants, hourlies



# InvertNet Collaborating Curators

- A. Cognato, MSU
- G. Courtney, J. VanDyk, ISU
- J. Holland, Purdue
- R. Holzenthal, P. Tinerella, Minnesota
- P. Johnson, SDSU
- H. Klompen, M. Daly, OSU
- J. Rawlins, R. Davidson, J. Fetzner, Carnegie Museum
- D. Rider, G. Fauske, NDSU
- A. Short, Kansas
- R. Sites, Missouri
- D. Young, Wisconsin-Madison
- J. Zaspel, Wisconsin-Oshkosh
- G. Zolnerowich, KSU
- D. Rubinoff, U Hawaii
- T. Roberts, U Iowa

