THE OHIO STATE UNIVERSITY

Luciana Musetti,
Sara Hemly,
Matt Elder &
Norman Johnson

Present:

The good, the bad and the ugly: The wild, wild west of specimen databasing

Triplehorn

Introduction

Digital capture of data in insect collections largely lags behind comparable efforts for plants and vertebrates. This is often attributed to the much larger scale of entomological collections. Unfortunately, few data are available to critically assess the bottlenecks in the digitization process so as to develop new tools and procedures to address this lag.

We have previously reported on the costs in time and money for curation and digitization for a single taxon. Here we compare two large sets of specimens: a general collection of **Carabidae** and a specialist collection of **Tenebrionidae** (Coleoptera). We anticipated that the different characteristics of these two types of collections would result in significant disparities in the time needed for digitization.

Methods

We recorded the time needed for each step in the digitization process, including adding unique identifiers to specimens, label data transcription, and locality georeferencing.

Specimen occurrence and taxonomic data are managed within the **xBio:D** platform and the underlying OJ Break API. The database engine is an Oracle RDBMS implementing the Association of Systematic Collections information model. In addition, this platform supports data from published literature, media, and characters (morphological and molecular, quantitative and qualitative).

For more information on the **xBio:D platform** or to explore possible collaboration, please see *xbiod.osu.edu/osucWiki/Main_Page* or contact *Johnson.2@osu.edu*.

About xBio:D

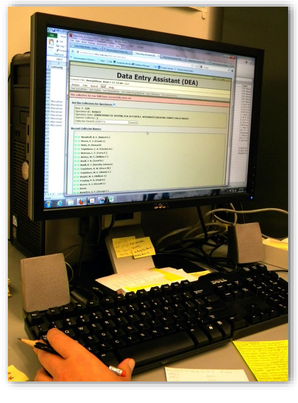
Workflow Description

1) Check & Update Taxonomic Names – check catalogs and/or the original literature. Not a mandatory step, but one that we think is necessary and that we adopted.

2) Specimen Handling

a) Add Unique ID – add a plastic tag containing a barcode label and the specimen unique ID in human readable format.

b) Transcribe Specimen Label – copy all the label data into MS Excel Data Entry template



3) Georeferencing – look up geographic coordinates using various online sources.

The three categories described above occur independently, coming together at the final step of data upload.

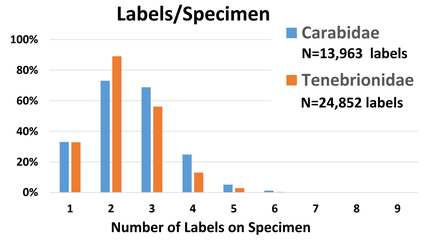
For the purpose of this analysis, **Total Databasing Time** is defined as the sum of the **steps 2 and 3**. The process of checking and updating taxon names (step 1) is not continuous, and we were unsuccessful in effectively quantifying the effort involved in that task.

A fourth step in our workflow is **data upload**. We use a web-based application, the **Data Entry Assistant (DEA),** to automate the process and add another level of quality assurance. In the course of this project, we moved to a more efficient and much faster version of the DEA. As a result the upload times for the two datasets are not comparable and are not included in this report.

Dataset Comparisons

Number of Labels

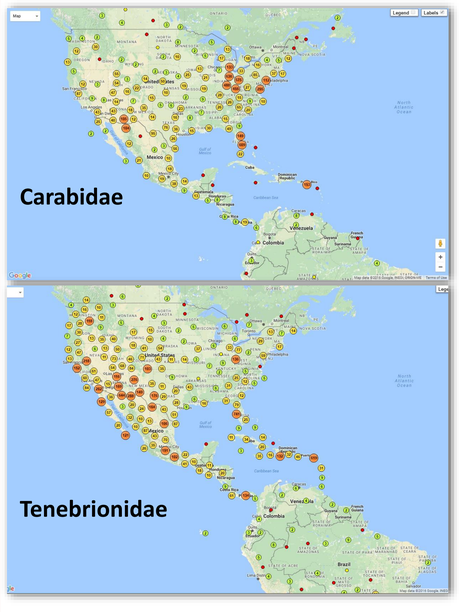
Data transcription is made more difficult when the specimen has multiple labels, perhaps printed on both sides or folded. There were 2.06 labels/specimen (Max. 9 labels) in Carabidae versus 1.94 labels/specimen (Max. 7 labels) in Tenebrionidae.

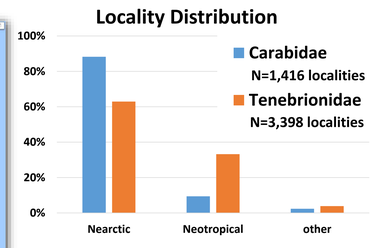


Multiple specimen labels

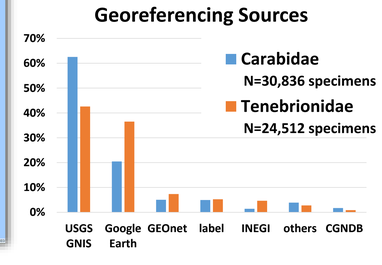
Localities

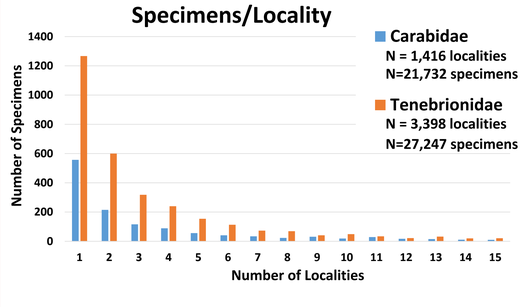
The tenebrionid dataset **(33.2%)** has a **higher percentage of specimens collected south of the United States** than the carabid dataset **(9.4%)**.





Georeferencing localities from outside the U.S.A. was **more laborious,** frequently demanding use of >1 source.





Of the **4,392** localities georeferenced, only **422 (9.6%) are shared between the two datasets.**

On average, the tenebrionid dataset **(8.0 specimens/ locality)** has **fewer specimens from each locality recorded** than the carabid dataset **(15.3 specimen/locality)**.

We also examined other variables such collectors, determiners, and presence of genitalia vials attached to the pin, but the differences observed were negligible.

Time Budget

DATABASING TASK TIME BUDGET – SPECIMENS/HOUR			
	Carabidae	Tenebrionidae	% Difference
Total # spms	21,732	27,247	
Adding identifiers	137.81	132.40	4.0%
Label transcription	87.78	91.99	4.7%
Georeferencing	163.73	114.29	35.6%
Total	**40.39**	**36.80**	9.3%

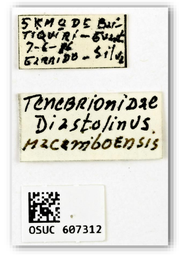
Georeferencing (specimen per hour) was **35.6% more time-intensive** in Tenebrionidae.

Despite that, the **Total Databasing Time** between the two datasets **differed by only 9.3%**.

Label transcription was the most time-consuming part of the digitization process for both datasets. Number of labels (Carabidae) & difficult to interpret labels are contributing factors.

DATABASING TASK TIME BUDGET – MINUTES/SPECIMEN				
	minutes/specimen		% of total time	
	Carabidae	Tenebrionidae	Carabidae	Tenebrionidae
Total # spms	21,732	27,247		
Adding identifiers	0.44	0.45	29.3%	27.8%
Label transcription	0.68	0.65	46.0%	40.0%
Georeferencing	0.37	0.52	24.7%	32.2%
Total	**1.49**	**1.63**		

Difficult labels



Georeferencing is particularly troublesome outside of the United States and Canada, and remains a major barrier. **A community-sourced clearinghouse of standardized georeferenced localities is needed to maximize the benefits of digitization and minimize redundant and variable work between institutions.**

Often unappreciated, though, is the difficulty in determining the current specialist opinion on **taxonomic names and their status**. Despite widespread efforts of data aggregators, **the low priority of taxonomic cataloging among funding agencies and even the scientific community is a major roadblock to a true catalog of life.**

The Tagline

1) **GOOD**: The two datasets differ in many dimensions, and these end up cancelling out each other in terms of the overall databasing effort. This suggests that budget calculations can be effectively based on a generalized time budget.

2) **BAD**: **Georeferencing** has long been recognized as a **roadblock**, and this obstacle remains.

3) **UGLY**: The lack of **comprehensive** and **easily accessible** taxonomic authority files.

Access our Specimen Data:

- **Local web portal**: *hol.osu.edu*
- **Global Biodiversity Information Facility** (GBIF): *www.gbif.org*
- **Symbiota Collections of Arthropods Network** (SCAN): *scan1.acis.ufl.edu/*

Credits:

- To Zach Hurley, for good work & friendship
- To the many **undergraduate curatorial assistants** who have done the bulk of the data entry work described herein. We would have accomplished very little without them.

Partial support for this work was provided by the **National Science Foundation in the form of a grant** No. DEB 60047322 *Digitization PEN: Integration of data from Triplehorn Insect Collection with the Southwestern Collections of Arthropods Network* to Johnson & Musetti. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. **ECN & ICE - September, 2016.**