Caveat Consumptor Notitia Museo: 
Let the Museum Data User Beware

RUNNING TITLE: Museum Collection Biases and Errors

ABSTRACT

Issue: Lot accession information from natural history collections represents a potentially vital source of large datasets to test biodiversity, biogeography and macroecology hypotheses. But does such information provide an accurate portrayal of the natural world? We review the many types of bias and error intrinsic to museum collection data and consider how these factors may affect their ability to accurately test ecological hypotheses.

Evidence: We considered all Texas land snail collections from the two major repositories in the state and compared them with an ecological sample drawn across the same landscape. We found that museum collection localities were biased in favor of regions with higher human population densities and iconic destinations. They also tended to be made during attractive temporal windows. Small, uncharismatic taxa tended to be under-collected while larger, charismatic species were over-collected. As a result, for most species it was impossible to use museum lot frequency to accurately predict frequency and abundance in an ecological sample. Species misidentification rate was approximately 20% while 4% of lots represented more than one species. Errors were spread across the entire shell size spectrum and were present in 75% of taxonomic families. Contingency table analysis documented significant dependence of both misidentification and mixed lot rates to shell size and family richness.

Conclusion: Researchers should limit their use of museum record data to situations where their inherent biases and errors are irrelevant, rectifiable, or explicitly considered. At the same time museums should begin incorporating expert specimen verification into their digitization programs.

KEYWORDS: Sampling bias, specimen labeling error, misidentification, data mining, museum records, body size
INTRODUCTION

Over the last two decades natural history collections have invested considerable effort into digitizing collection and accession / lot information. Aided by funding from the European Commission (e.g. the Europeana Digital Collection Portal) and National Science Foundation (e.g. Advancing Digitization of Biological Collections program) remotely accessible digital museum record data are now being produced at an industrial scale (Blagoderov et al. 2012). Museums are actively promoting their use (Krishtalka & Humphrey 2000, Funk 2018) as an important resource to test a wide range of biodiversity hypotheses (Graham et al. 2004, Suarez & Tsutsui 2004, Lavoie 2013). The macroecological community has responded by beginning to mine and utilize such information to test hypotheses (e.g. Babin-Fenske et al. 2008, Foley et al. 2008, Ramirez-Villegas et al. 2012, Economo et al. 2018).

It is unquestioned that museum collections represent an invaluable data source for research into systematics (Wen et al. 2015), evolution (Avila-Arcos et al. 2012), biological conservation (Drew 2011) and environmental response (Pyke & Ehrlich 2010). The idiosyncratic nature of museum accessions (Ponder et al. 2001), however, raises a larger question about their ability to accurately reflect ecological and biogeographic pattern. Does bias and error in these data fundamentally limit their utility for drawing accurate conclusions? Or are they simply a nuisance which only adds unbiased noise into a dataset?

We approach these questions as both generators and consumers of museum collection data across multiple taxonomic groups. We know from first-hand experience not only their potential power, but also the many ways in which their naive use can lead to erroneous conclusions. We start by reviewing the many ways in which bias and error can creep into museum record databases, providing examples based on our experiences with land snail, vascular plant, crustacean, and lepidopteran collections. We then move beyond such anecdotal examples to empirically analyze these issues using Texas land snail museum records. We end by considering the limitations that end users should place on the use of unverified, digitized collection data and possible remediations available to the natural history collections community.

POTENTIAL SCOPE OF THE PROBLEM

Minimally three steps lie between an ecological pattern in the field and its representation as museum record data: collection, labeling, and curation. Each is associated with potential biases or errors (Soberón et al. 2000, Newbold 2010):
(1) Collection Bias. Many collections were often never intended to represent an unbiased ecological sample of the natural world: rather they explicitly represent the particular interests (and aversions) of their collectors. For instance, collection sites may be concentrated near collectors’ homes and home institutions (Palmer 1995), proximate to easily accessible roads and trails (Soberón et al. 2000), or focused on exotic or iconic landscape locations (Ponder et al. 2001). We have also seen individuals recollect the same area over and over again explicitly because of the site’s high productivity, thereby avoiding the risk of a return from the field with little to show for their efforts.

There are additional facets to the decidedly biased nature of how a collector collects:

First, collecting may occur more frequently on some dates. For instance, while an undergraduate assisting his collegiate herbarium JCN noticed with amusement how many sheets for the charismatic prairie fringed orchid (Platanthera praecella Sheviak & M.L.Bowles) in Iowa were made on July 4: the USA Independence Day holiday. Certain dates may also be avoided – for instance mid-April to mid-May when the North American academic schedule is filled with exams and grading. If an animal’s active or identifiable life stage occurs during that period, its distribution may be underrepresented. Perhaps it is thus not surprising that the land snail Vertigo alabamensis Clapp – which is present as identifiable adult shells only during April and May – was long believed present from only six Alabama sites when in fact it is characteristic of coastal pine forests across 1500 km extant (Nekola & Coles 2010). Similarly, the grassland skipper Atrytonopsis hianna (Scudder) – which flies only for a brief window from late April to early May – had not been collected in eastern Iowa for over a century even though it is actually present on many hill prairies (Schlicht et al. 2007).

The taxonomic focus of many collectors also means that their collections often emphasize favorite groups while ignoring others. Additionally, charismatic species with a body size large enough to be easily observed but not so big as to make documentation problematic (e.g., large herbaceous plants like Amorphophallus or Musa) may be preferentially collected. For instance, many North American groundwater crustacean collections lack dominant ostracod and copepod groups because field work is often done without magnification and with the most commonly used nets having mesh sizes too large for organism retention. Collection rates may also be inversely correlated to species abundance. For instance, there were more collections of rare lady’s slipper orchids (Cypripedium spp.) in JCN’s undergraduate herbarium than weedy
dandelions (*Taraxacum* spp.). And, in the Iowa Lepidoptera collection database assembled by Schlicht et al. (2007) there are roughly as many records for the endangered prairie obligate *Hesperia ottoe* Edwards and fen specialist *Euphyes conspicua* (Edwards) as for *Nymphalis antiopa* (Linnaeus) and *Polygonia comma* (Harris) which likely occurs in every woodlot in the state. This bias is at least partially promoted by field survey funders (e.g. USA State Wildlife Grant programs) who often explicitly target rare species as opposed to whole community data.

**(2) Labeling errors.** Data reported on an accession label may not be accurate. First, material may be taxonomically misassigned. This has been recognized as an important source of error since the dawn of natural history collection record digitization (Crovello 1967). In some cases material was correctly identified at the time of accession but the current name differs. In others the species to which a specimen was correctly assigned was later split with the specimen now representing one of the newly described taxa. And sometimes material was simply misidentified even using taxonomic standards of the time. For instance, all Ontario lots of *Vertigo clappi* Brooks & Hunt in the Museum of Comparative Zoology, Royal Ontario Museum, and University of Michigan Museum of Zoology were incorrectly identified as *V. milium* (Gould) by collector John Oughton (Nekola 2009). Additionally, for taxonomic groups that can possess more than one individual per lot (for instance land snails, aquatic crustaceans and small plants), multiple species may be represented. Such mixed lots may have labels which correctly identify one of the included individuals, or none – with misidentifications representing all of the above error types.

Errors can also exist in site information. Sometimes the collector reported a location which simply does not exist. For instance, the type location for the land snail *Oreohelix socorroensis* Pilsbry was identified by W.D. Hartman as being from the ‘Negra Mountains, Socorro County, NM’ (Pilsbry 1948) even though such a mountain range does not exist in that county or anywhere else in the state. And, the type locality for the groundwater crustacean *Texanobathynella bowmani* Deboutteville was reported from Roaring Springs, ‘Travies Country’, Texas (Deboutteville et al. 1975). No such county (or country, for that matter!) exists. While there is a Travis County, no Roaring Springs occurs there. And while a Roaring Springs does exist in Motley County, this site was explicitly stated to not support this species. Sometimes collection localities were misreported: the land snail *Ashmunella levettei* (Bland), for example, was originally reported by collector G.M. Levette from Santa Fe, New Mexico when in fact it
had been obtained from the Huachuca Mountains of southeastern Arizona (Pilsbry 1948). And sometimes political or geographic place names have changed: older collections from El Paso County, Texas may now actually reside in Hudspeth County due to boundary changes in 1917.

Collectors also sometimes purposefully report incomplete or inaccurate location information. For instance, the type locality for the land snail *Haplotrema costatum* Smith was simply reported as “Cave 12-19, Tulare County, California” to keep the location secret from casual speleologists (Smith 1957). Because the field notebooks of the collector, Raymond deSaussure, have subsequently been lost there is now no way to locate this site. Sometimes more nefarious reasons exist: California professional collector A.W. Crawford purposely misstated the type location for *Monadenia circumcarinata* (Stearns) twice – first as ‘Turlock, Stanislaus County’ and then, after being challenged about its impossibility, as being from ‘near Columbia, Tuolumne County’ (Hanna & Smith 1954). He did this to maintain a monopoly on shells entering the market and to hide the location of an associated mining claim (Barry Roth, personal communication). It took 75 years for the actual location to be stumbled upon – over 20 km from the second claimed locality and in a different drainage system!

Lastly, labels themselves may be illegible or incomplete. The latter issue seems especially common in terms of localized geography, habitat, collection date and collector for older lots, and exists because at the time of accession this information was considered unimportant. This exact problem was bemoaned over a century ago by none other than Alfred Russel Wallace (Slotten 2004).

(3) Curation Errors. A final suite of potential problems is associated with activities following accession. While most relate to specimen or data loss through incorrect labeling media or storage conditions (e.g. Bynes Disease; Tennent & Baird 1985), the misassignment of labels with specimens can also occur and is often difficult to rectify. For instance, the label for Los Angeles County Museum lot 93636 states Santa Catalina Island, California as the locality. While it does contain endemic *Micrarionta* from that island, mixed in are a few *Oreohelix* representing a race from Idaho, Oregon, or Washington. At some point these specimens must have been removed from their correct lot and mistakenly returned to this one.
MOUNTAINS OR MOLEHILLS? HOW IMPORTANT IS COLLECTION BIAS AND
LABELLING ERROR?

How important are these potential biases and errors? Are they just uncorrelated noise
that simply represents a statistical nuisance? Are they swamped out by accurate data as
suggested by Lavoie (2013) who pointed out that georeferencing error in herbarium labels –
leading to 16% of sample locals being placed in the ocean – alternatively means that the other
84% could be correct. Or, perhaps they fundamentally alter distributions leading to an inaccurate
portrayal of the original pattern.

To move beyond the above anecdotes we attempt to empirically document how much
these issues alter statistical distributions and potentially inhibit our ability to accurately
document biogeographic and ecological pattern. We do this by considering Texas land snail
records. Terrestrial gastropods represent an ideal system to address these issues: with over
35,000 species believed to exist globally they represent the second most diverse molluscan group
(Barker 2001). At the same time their taxonomy is relatively mature, with the description of new
species having slowed in North America to less than 1% of the total fauna per decade (Nekola
2014). Both continental (e.g. Pilsbry 1948) and regional (e.g. Baker 1939, Cheatum et al. 1974)
identification resources extend back over 70 years and arguably represent the earliest whole-
fauna treatments produced for any non-lepidopteran invertebrate group. Additionally, molluscan
holdings in the United States and Canada museums are second only to insects in total number of
recorded specimens and represent the best-sampled class of metazoans on a per-species basis
(Sierwald et al. 2018).

We consider the Texas fauna because, as part of a larger project to assess the
conservation status of terrestrial gastropod species within the state, we made on-site visits to both
major holdings: the Perot Museum of Natural History and University of Texas El Paso (UTEP)
Centennial Museum. Since they were assembled by the foremost Texas land snail experts during
the latter half of the 20th Century (Elmer P. Cheatum, Richard W. Fullington, William J. Platt
and Artie L. Metcalf), we assumed these repositories would be relatively well curated with
minimal misidentification due to simple lack of familiarity with the taxa.

Datasets

Museum Record Data. We entered label data and verified identifications for all Texas-
sourced lots, representing 2582 from the Perot and 1190 from the UTEP collections. Slugs and
all members of Succineidae (save *Succinea luteola* Gould) were excluded because they could not be accurately assigned to species based on shell / internal plate features. Drift, fossil or subfossil material was also excluded because our focus was on the conservation status of the extant fauna. In total 150 species were represented out of 211 reported from the state (Supplementary Material 1).

**Ecological Data.** Comparative ecological field data were collected by JCN, and represent 97 sites spread across the state, including most major habitat types, 122 species, 1048 lots and 52,065 individuals (Supplementary Material 2). Sites represented high-quality examples of each community type. Anthropogenic habitats were excluded while some sites were selected because of their known presence of rare species populations. Each selected 100-1000 m² area was sampled through hand collection of larger shells and litter collection for smaller taxa. This approach provides the most complete assessment of site faunas (Oggier et al. 1998, Cameron & Pokryszko 2005). Litter sampling was made at places of high micro-mollusc density such as loosely compacted leaf litter lying on top of highly compacted damp mineral soil or humus (Emberton et al. 1996). This was processed in the field using a shallow sieve of 2 mm mesh nesting loosely inside a sieve of 0.6 mm mesh. The ~500 ml of retained litter per site was slowly and completely dried in the laboratory and then passed through a standard sieve series (ASTME 3/8" (9.5 mm), #10 (2.0 mm), #20 (0.85), and #40 (0.425 mm) mesh screens). Sorted fractions were hand picked against a neutral-brown background with all shells and shell fragments being removed and assigned to species (or subspecies). The total numbers of shells per species per site were recorded, as were the number of unassignable immature individuals and fragmentary shells.

**Other data.** Maximum shell dimension for each species was based on Nekola (2014). In addition, the North American species richness of families was determined using Nekola (2014) in conjunction with Bouchet & Rocroi (2005) and Schileyko (2006). County-scale demographic and geographic data was obtained via the Texas Open Data Portal (https://data.texas.gov/).

**Issue 1: Collection Biases**

**Geography.** Mapping of Perot and UTEP lots per Texas county shows they are spread unevenly across the landscape (Figure 1). While this type of bias has often been portrayed as being idiosyncratic (Ponder et al. 2001), the fact that the same counties tend to be most represented in both collections implies that predictive factors are at play. We conducted a multiple linear regression to determine how much variation in total number of lots per county
could be explained by log-transformed human county population (based on the 1980 USA Census), log-transformed county land area (discounting water bodies) and a dummy-variable representing the presence/absence of a USA National Park (Big Bend or Guadalupe Mountains). Log-transformation was used because it provided the best linear model fit. No interaction terms were found significant. Each of the three factors significantly (p<0.00815) predicted lot numbers per county and in total explained almost 30% of observed variance (Figure 2).

Because the two repositories are separated by 1000 km, the impact of distance could only be investigated separately. We first calculated the distance from each museum to each county centroid. We then saved residuals from the above multiple regression model run separately for each repository and determined the significance of distance-to-repository in these residuals. We again used log-transformed data because it provided the best linear model fit. While distance to UTEP was a highly significant additional predictor (p<0.0000001), explaining over 25% of remaining variance (Figure 3), distance to Perot Museum was not (p=0.624). However, a trend of decreasing residual scores with increasing distance was apparent up to about 750 km. When these more remote counties were excluded, distance was found to be a significant predictor (p=0.000723), explaining an additional 3% of residual variation. It appears as if the attractive force of Trans-Pecos mountains was able to overcome whatever natural inclination existed in Perot Museum researchers to collect close to home.

**Collection Date.** We could not empirically test for this bias because: (1) only 950 records (~25%) provided collection date beyond year; (2) there was no consistent pattern regarding whether dates were provided as day/month/year or month/day/year. As a result we had no way of knowing collection date when day and month was <=12. However, some insights can be gleaned from comparable data assembled by Nekola (2009) for all Ontario and Quebec land snail collections of conservation importance held in the Academy of Natural Sciences of Philadelphia, Carnegie Museum, Museum of Comparative Zoology, National Museum of Canada, Royal Ontario Museum, and the University of Michigan Museum of Zoology. A log-likelihood contingency table test (Sokal & Rohlf 1981) demonstrates that for the 75% of records collected during the growing season, May (204) and August (168) had significantly (p=0.003; test statistic = 13.97 on 3 df) more accessions than June (133) and July (118). It seems at least possible that May is more represented because Canadian collectors were anxious to get out in the
field after a long winter, with August being more represented because they were anxious to get
into the field one last time before the close of the growing season.

**Body Size.** This potential bias was considered by comparing the distribution of maximum
shell dimension for all museum lots to that from species occurrences (e.g. lots) and individuals
encountered via ecological sampling. Both log-likelihood contingency table ($p<0.000000001$)
and Kolmogorov–Smirnov ($p<0.000009$) tests demonstrated (Figure 4) that museum lots
significantly underrepresented the frequency of small and over-represented the frequency of
large species as compared to both occurrences and total individuals encountered during
ecological sampling.

**Relative Species Abundance.** This potential bias was investigated thorough Spearman-
Rank Correlation on the total number of museum lots for each species vs. total number of lots
and individuals encountered through ecological sampling. Only the 98 species present in both
datasets were used. While the overall correlation was significant and positive (Figure 5) for
museum lots vs. ecological occurrences ($p=0.0052; \rho=0.28$) and vs. ecological abundances
($p=0.00052; \rho=0.34$), visual inspection of the data suggested that this might be driven by only
the most common species. To investigate this we repeated Spearman-Rank correlation
following removal of the most frequent species from the museum records, then again after
removal of the first and second most abundant, first to third most abundant etc. to removal of all
but the two rarest species. Spearman $p$-values were recorded and plotted against the rank order
of which species were retained. These analyses did indeed show that relationship significance
was limited to only those data which included the five (vs. ecological lots) or eighteen (vs.
ecological abundance) most frequent species in museum collections (Figure 5). While at some
arbitrarily small number of species a non-significant Spearman-Rank correlation would be
expected, we find it non-trivial that there was no significant correlation between number of
museum lots and ecological occurrences for the least common 93% of species, and between
museum lots and total encountered ecological individuals for the least frequent 81% of species.

**Issue 2: Labeling Errors**

Given the effort that has been made to identify and correct geolocation / place name
errors in museum record data (e.g. Gulanick et al. 2007), we chose to concentrate instead on the
less investigated issues of lot misidentification and mixing. To do this we encoded each lot as
representing either a correct or incorrect initial determination and as being either mixed ($n>1$
species) or unmixed (n=1 species). For purposes of this analysis, labels with misspellings and/or outdated nomenclature were noted but not considered misidentified. In general, KEP and BTH verified snails with maximum shell dimensions >= 10 mm, while JCN was responsible for all smaller taxa. In cases where a lot could not be confidently allocated, consensus across all three investigators was used to assign a species name. For mixed lots, both the identity of principle species (as indicated on the label) and all other species in that sample were recorded.

The frequency distribution of misidentified and mixed lots across all species was illustrated using Cumulative Rank Frequency Distribution (CRFD) plots in log-log space (Newman 2005). In these plots the upper-most left point represents to total number of species with at least one misidentified lot, the next point to the right represents the total number of species with at least two misidentified lots, the third the total number of species with at least three misidentified lots and etc. until the final point in the lower right corner which represents the single species with the highest number of misidentified lots. Likelihood ratio contingency table tests were used to test for non-independence in misidentification and mixed lot rates vs. repository, body size and family richness. To do this, maximum shell dimension was converted into five logarithmically increasing size classes: <2.5 mm; 2.5-<5 mm; 5-<10 mm; 10-<20 mm; and 20+ mm. North American family richness was converted into eight octave-binned categories following Preston (1948): 1 species; 2-3 species; 4-7 species; 8-15 species; 16-31 species; 32-63 species; 64-127 species; and 128-255 species. The change in body size distribution between initial vs. verified identifications and between principle and additional shells in mixed lots was also documented through use of log likelihood ratio contingency table tests.

Lot Misidentification. Over 20% of lots (769 out of the 3772) were misidentified at the species level across both collections, while 145 (3.8%) were misidentified at the genus level. This led to some critical errors: for instance, the Neck (1980) report of a 700 km range extension for the rare *Gastrocopta riograndensis* (Pilsbry & Vanatta) into the Trans-Pecos is based on misidentification of the common *Gastrocopta pellucida* (Pfeiffer). Sadly, this error was passed on in the range maps of Nekola & Coles (2010) because it was assumed that the published record was accurate. Additionally, all Texas reports for the exotic *Opeas pyrgula* Schmacker & Boettger and *Subulina octonea* Bruguire in Cheatum et al. (1974) represent misidentifications of *Lamellaxis gracilis* (Hutton). The CRFD of misidentification rates per species (Figure 6) illustrates a negative power law shape with almost 50% of all misidentifications being associated
with five taxa \((\textit{Glyphyalinia umbilicata} \text{ (Singley in Cockerell), \textit{Hawaiiia miniscula} (A. Binney), \textit{Linisa texasiana} (Moricand), \textit{Gastrocopta sterkiana} Pilsbry, \textit{Helicina orbiculata tropica} Say})\) and 72 species (46% of total) having no misidentifications. In spite of this, misidentifications were spread across 24 families (75% of total). Even though correctly identified, outdated generic and species-level nomenclature were observed in 921 (24.4%) and 78 lots (2%), respectively. Contingency table analysis illustrated no dependence \((p=0.124)\) in misidentification rate between the two repositories (Table 1A). However, there was strong dependence \((p<0.00000001)\) in body size with highest misidentification rates (21-27%) occurring in the three intermediate size classes (Table 1B). Strong dependence \((p<0.00000001)\) was also noted in family richness, being lowest (0% misidentified) for monotypic families and highest (36.1%) in families possessing 32-127 species (Table 1C). Lastly, misidentification led to a significantly \((p<0.00000001)\) more uniform shell size distribution (Table 1D).

**Lot Mixing.** Almost 4% of lots (149 out of the 3772) possessed multiple species, with a total of 196 additional occurrences being recorded. Again, this led to some critical errors with two species – \textit{Gastrocopta rogersensis} Nekola & Coles and \textit{Gastrocopta similis} (Sterki) – having been previously left out of the state fauna. The CRFD of mixed lot incidence (Figure 7) again illustrates a negative power law shape with almost 50% being associated with four species \((\textit{Hawaiiia miniscula, Gastrocopta contracta} \text{ (Say), Gastrocopta procera} \text{ (Gould), Gastrocopta pellucida})\), and 120 species (over 75% of total) possessing no mixed lots. The CRFD for additional mixed-lot species records displayed a similar shape with five \((\textit{Gastrocopta sterkiana, Gastrocopta pentodon, Hawaiiia miniscula, Gastrocopta pellucida, Helicodiscus singleyanus})\) representing over 42% of all material and 106 species (over 2/3 of the total) being unrepresented. Mixed lot rate was more than double in the Perot vs. UTEP collection \((p=0.00000053; \text{ Table 2A})\), and was also greatly influenced \((p<0.00000001)\) by shell size with the lowest rates (~1%) occurring at the largest and highest rates occurring (9%) in the smallest size class (Table 2B). Strong dependence \((p<0.00000001)\) was also noted in log-binned family richness, with error rates being lowest (0%) in families with three or fewer species and highest (8.5%) in those possessing 64-127 species (Table 2C). Lastly, there was no difference \((p=0.50)\) in shell size distribution of principle vs. additional shells in mixed lots (Table 2D).

**Erroneous and missing location data.** While we chose to not explicitly investigate rates of misspelled / erroneous geographic place names or incorrect georeferencing we note that this
type of error was common. Also of concern were instances where label information was incorrectly transcribed from collection notebooks. Perhaps the most egregious example were two dozen UTEP Hutchinson County lots collected on June 19, 1971 either from “along road to Adobe Wells” or “along creek 1 mile NE of Adobe Wells Monument”. No such locality exists, although there is a historical site called “Adobe Walls” in the county. Among these accessions were eleven species either unknown from Texas or restricted to the highest peaks in the Trans-Pecos. Their presence in the low, arid plains of the Texas Panhandle would have been remarkable. However, other UTEP lots show that immediately preceding this date a large number of samples had been made in the northern New Mexico mountains where all these species commonly occur. Clearly these putative Hutchinson County records were collected in New Mexico, and if taken at face-value would have greatly altered range and conservation status for almost 5% of the entire Texas land snail fauna. We also note that some dozens of R.W. Fullington’s Perot Museum lots from Guadalupe Mountains National Park only provide site collection numbers. Because his collection notebook is missing their actual location cannot be determined.

**IMPLICATIONS FOR MUSEUM RECORD USE**

Bias and error in museum record data represent more than just a few anecdotal examples that simply add uncorrelated noise into relationships. Our analyses suggest that at least a dozen issues significantly impact the representativeness of museum data in the testing of biodiversity and ecology hypotheses (Table 3). We found collection localities to be biased in favor of regions with bigger human populations or iconic destinations. We saw that collections tended to be made during attractive temporal windows, and nearer to the repository location. We observed that small, uncharismatic taxa tended to be avoided while larger, charismatic species were favored. As a result, for most species within the fauna it was impossible to predict frequency and abundance in an ecological sample from the number of museum lots.

Besides these biases, errors rates were high and had profound consequences on accurate documentation of biogeographic and ecological pattern. For instance, the 20% identification error rate led to a significant flattening of the apparent body size distribution. These misidentifications were most frequent in species of intermediate body size and from the most diverse families. Almost four percent of lots were also found to represent more than one species,
with mixing rate being higher in taxonomically diverse families and in species with smaller body size. This led to underreporting of species richness from site to regional scales. In addition we saw label information which had been incorrectly transcribed from collection notebooks leading to wildly unlikely occurrence records. It is sobering to realize that a field note transcription error from a single collector on a single day led to the potential misreporting of range and habitat for 5% of the state fauna. And we noted labels which provided incomplete data because the collection notebooks they referenced have been lost.

Many of these issues appear general in nature. Similar non-representative locations have been noted for other molluscan (Ponder et al. 2001) and vascular plant (Palmer 1995) collection data. The misidentification rate we observed is also well within the bounds reported for other taxonomic groups, including 27% for freshwater clams (Shea et al. 2011), 20% for bird holdings in the Šariš Museum (Mikula et al. 2018), 20% for fungi used in molecular analyses (Bridge et al. 2003), and 10% for germplasm holdings of *Citrullus* (watermelons) in international seed banks (Guzzon & Ardenghi 2017). Some groups are even more impacted such as an 83% error rate in an online database of *Euscelidia* robber-flies (Meier & Dikow 2004) and 56% rate for herbarium accessions in the flowering plant genus *Aframomum* (Goodwin et al. 2015). Such misidentification errors have led to profound confusion in phylogenetic tree interpretation (Ó Foighil et al. 2009).

In fairness we should point out that some of these issues are not unique to museum records in general or to these repositories in particular. Non-representativeness in sample location also exists in our comparative ecological database where sampling was limited to high quality natural habitats. As a result anthropogenic habitats that make up a considerable proportion of the landscape were excluded. Reported frequencies and abundances are thus strongly biased against anthropophilic species. We also acknowledge that ecological sample sites are clustered within the landscape, in large part because of spatial autocorrelation of habitat occurrences in relation to topography, environment, and land use patterns. Such non-independence is simply a reality for any study involving natural habitats. We also have no illusions that future researchers will agree with all our identifications. For instance, JCN found a 4% error rate in his personal holdings of the land snail genus *Euconulus* following DNA sequence analysis that allowed documentation of those shell features which provide accurate species-level diagnoses (Horsáková et al. 2019). Without such additional information, however,
misidentification based on visual cues must be viewed as a perhaps irresolvable component of ecological and conservation datasets.

It is thus essential for researchers who use museum data to demonstrate that these biases and errors do not interfere with accurate pattern documentation and hypothesis testing. For instance, while we are not overly concerned about the systematic undersampling of anthropophilic species in the ecological dataset as it was collected to document natural community structure, the bias of museum collections in favor of areas with denser human populations could well confound analyses and lead to incorrect interpretations.

**REMEDIATION PROSPECTS**

What – if anything – can be done to minimize these problems? Because collections already in museum holdings likely were never collected with the intent of accurately documenting the abundance and distribution of all species within a fauna, the only avenue for potential remediation lies in algorithms that correct for such biases. However even a very simple mathematical transformation can generate artifacts that fundamentally alter pattern (Nekola et al. 2008). Who can guess what phantoms might be created through even more complicated calculations? Furthermore, we observed considerable idiosyncrasy in bias and error rates between collectors and repositories (e.g., the stronger distance-to-repository effects at UTEP and higher mixed lot rate at Perot). As a result, creation of a single generic model to correct for museum record bias and error appears unlikely.

Making museum data better suited for testing general ecological hypotheses will require that future collecting efforts: (1) include undersampled regions, time periods, body-size classes, taxa-groups, etc., and (2) be made in accordance with accepted ecological sampling protocols. This highly structured sampling, however, may be unappealing to more taxonomically inclined collectors, and with there being little or no existing money in institutional budgets to accommodate such activities, their prioritization could represent a hard sell to funders and administrators.

Dealing with label error rates may represent a more tractable problem. Issues with typographical errors, outdated taxonomy, geographical assignments and poor geolocation can potentially be resolved using computer code (Guralnick et al. 2007). However, this cannot fix lot misidentification and mixing errors whose correction requires the on-site presence of a well-
trained human mind. Sadly, most ‘industrial-scale’ efforts to digitize museum data appear solely reliant upon robotics (e.g., Bertone et al. 2012, Blagoderov et al. 2012, Tegelberg et al. 2012) with there being little or no effort made to ensure that captured data are correctly identified. Clearly robust Quality Assurance/Quality Control procedures must be incorporated into the workflow of all future collections data modernization efforts.

Does this mean that online museum records are ‘junk data’ (Vilgalys 2003) to be avoided when testing ecological and biological hypotheses? Not at all! As pointed out by Pyke & Ehrlich (2010), museum collections shine in their ability to document change within populations over time. The issues we detail here generally do not apply to such studies – as long as labels provide accurate collection locations, dates and identifications. Such vetted museum records can also provide unique and invaluable insights into abundance and trait trajectories across entire groups of species (Ball-Damerow et al. 2014), and accurate estimates of environmental range (Phillips et al. 2006). Additionally, high-quality museum data are foundational to taxonomy, systematics and conservation planning (Drew 2011). Where we become concerned is when unvetted data are naïvely assumed to be accurately identified and representative of the natural world and then used to test large-scale biodiversity hypotheses. While some museum data may be appropriate to address some ecological questions, researchers must be cognizant of their inherent limitations and not ask too much – or the wrong things – of them.

A final comment: robotic mass-production of digitized museum record data needs to be carefully reevaluated. During a time when many taxonomic experts are being forced out of the field due to lack of employment, transitioning to highly automated data acquisition may actually be driving down the number of available jobs (McClain 2011). Yet, a potential win-win situation could exist if museums decide to leverage the demand for such big data into paid positions to validate identifications in conjunction with digitization. It must be made clear to administrators and funders that the limiting step in large-scale digitization efforts is not technology but rather the human capital required to ensure high quality data (Tulig et al. 2012).

REFERENCES


Pyke, G. H. & Ehrlich, P. R. (2010). Biological collections and ecological / environmental
research: a review, some observations and a look to the future. *Biological Reviews*, 85,
247–266.


Supplement 2, 1765-1906.

Iowa Press.

on mollusk collections in the U.S.A. and Canada. *American Malacological Bulletin*, 36,
177-214.

mussel species (Bivalvia:Unionidae): Contributing factors, management implications,
and potential solutions. *Journal of the North American Benthological Society*, 30, 446-
458.


Sciences*, 29, 21-46.

conservation purposes: an example using Mexican Papilionid and Pierid butterflies.
*Biodiversity and Conservation*, 9, 1441–66.


**DATA ACCESSIBILITY**: Because analyzed data include specific locality information for many protected species, and because access to private land was often granted only after agreeing to not publically broadcast findings, we are unable to make the full datasets upon which these analyses are based available to the general public. However, the corresponding author can share this data with individual researchers upon request.
**Table 1:** Factors associated with misidentification

A. **Repository**

<table>
<thead>
<tr>
<th>Repository</th>
<th>Perot</th>
<th>UTEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>2038</td>
<td>965</td>
</tr>
<tr>
<td>Misidentified</td>
<td>544</td>
<td>225</td>
</tr>
<tr>
<td>Percent</td>
<td>21.1</td>
<td>18.9</td>
</tr>
</tbody>
</table>

Log-likelihood test Statistic = 2.369; df = 1; p-value = 0.1238007

B. **Maximum shell dimension**

<table>
<thead>
<tr>
<th>Maximum Dimension</th>
<th>&lt;2.5 mm</th>
<th>2.5-&lt;5 mm</th>
<th>5-&lt;10 mm</th>
<th>10-&lt;20 mm</th>
<th>20+ mm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>488</td>
<td>619</td>
<td>745</td>
<td>694</td>
<td>457</td>
</tr>
<tr>
<td>Misidentified</td>
<td>57</td>
<td>229</td>
<td>211</td>
<td>184</td>
<td>88</td>
</tr>
<tr>
<td>Percent wrong</td>
<td>10.5</td>
<td>27.0</td>
<td>22.1</td>
<td>21.0</td>
<td>16.2</td>
</tr>
</tbody>
</table>

Log-likelihood test statistic = 68.1257; df = 4; p-value <<0.000000001.

C. **Family richness within the North American fauna**

<table>
<thead>
<tr>
<th>Number of species</th>
<th>1</th>
<th>2-3</th>
<th>4-7</th>
<th>8-15</th>
<th>16-31</th>
<th>32-63</th>
<th>64-127</th>
<th>128-256</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>3</td>
<td>19</td>
<td>183</td>
<td>575</td>
<td>917</td>
<td>264</td>
<td>195</td>
<td>847</td>
</tr>
<tr>
<td>Misidentified</td>
<td>0</td>
<td>7</td>
<td>8</td>
<td>70</td>
<td>257</td>
<td>149</td>
<td>110</td>
<td>168</td>
</tr>
<tr>
<td>Percent wrong</td>
<td>0.0</td>
<td>26.9</td>
<td>4.2</td>
<td>10.9</td>
<td>21.9</td>
<td>36.1</td>
<td>36.1</td>
<td>16.6</td>
</tr>
</tbody>
</table>

Log-likelihood test statistic = 188.13; df = 7; p-value <<0.000000001

D. **Impact on body size distribution**

<table>
<thead>
<tr>
<th>Maximum Dimension</th>
<th>&lt;2.5 mm</th>
<th>2.5-&lt;5 mm</th>
<th>5-&lt;10 mm</th>
<th>10-&lt;20 mm</th>
<th>20+ mm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original ID</td>
<td>141</td>
<td>137</td>
<td>193</td>
<td>197</td>
<td>79</td>
</tr>
<tr>
<td>Verified ID</td>
<td>53</td>
<td>221</td>
<td>206</td>
<td>175</td>
<td>92</td>
</tr>
</tbody>
</table>

Log-likelihood test statistic = 64.0225; df = 4; p-value <<0.000000001
Table 2: Factors associated with lot mixing

A. Repository

<table>
<thead>
<tr>
<th>Repository</th>
<th>Perot</th>
<th>UTEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unmixed</td>
<td>2454</td>
<td>1169</td>
</tr>
<tr>
<td>Mixed</td>
<td>128</td>
<td>21</td>
</tr>
<tr>
<td>Percent</td>
<td>5.0</td>
<td>1.8</td>
</tr>
</tbody>
</table>

Log likelihood test statistic = 25.1626; df = 1; p-value = 0.00000053

B. Maximum shell dimension

<table>
<thead>
<tr>
<th>Maximum Dimension</th>
<th>&lt;2.5 mm</th>
<th>2.5-&lt;5 mm</th>
<th>5-&lt;10 mm</th>
<th>10-&lt;20 mm</th>
<th>20+ mm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unmixed</td>
<td>495</td>
<td>791</td>
<td>928</td>
<td>869</td>
<td>540</td>
</tr>
<tr>
<td>Mixed</td>
<td>50</td>
<td>57</td>
<td>28</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>Percent mixed</td>
<td>9.2</td>
<td>6.7</td>
<td>2.9</td>
<td>1.0</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Log likelihood test statistic = 92.889; df = 4; p-value <<0.000000001

C. Family richness within the North American fauna

<table>
<thead>
<tr>
<th>Number of Species</th>
<th>1</th>
<th>2-3</th>
<th>4-7</th>
<th>8-15</th>
<th>16-31</th>
<th>32-63</th>
<th>64-127</th>
<th>128-256</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unmixed</td>
<td>3</td>
<td>26</td>
<td>188</td>
<td>639</td>
<td>1099</td>
<td>396</td>
<td>279</td>
<td>993</td>
</tr>
<tr>
<td>Mixed</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>6</td>
<td>75</td>
<td>17</td>
<td>26</td>
<td>22</td>
</tr>
<tr>
<td>Percent mixed</td>
<td>0.0</td>
<td>0.0</td>
<td>1.6</td>
<td>0.9</td>
<td>6.4</td>
<td>4.1</td>
<td>8.5</td>
<td>2.3</td>
</tr>
</tbody>
</table>

Log likelihood test statistic = 63.300; df = 7; p-value <<0.000000001

D. Body size distribution of target vs. mixed species within a lot

<table>
<thead>
<tr>
<th>Maximum Dimension</th>
<th>&lt;2.5 mm</th>
<th>2.5-&lt;5 mm</th>
<th>5-&lt;10 mm</th>
<th>10-&lt;20 mm</th>
<th>20+ mm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td>50</td>
<td>57</td>
<td>28</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>Mixed</td>
<td>76</td>
<td>71</td>
<td>38</td>
<td>9</td>
<td>2</td>
</tr>
</tbody>
</table>

Log likelihood test statistic = 3.36457; df = 4; p-value = 0.4987669
Table 3: Summary of museum data issues and potential remediations

A: Collection Biases

<table>
<thead>
<tr>
<th>Type</th>
<th>Effect</th>
<th>Potential Remediation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Collection Biases</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collection location</td>
<td>Collections more likely from counties with greater human population, more land area, iconic destinations, and nearer to repository</td>
<td>Target new acquisitions from undersampled areas</td>
</tr>
<tr>
<td>Collection date</td>
<td>Uneven sampling across year / growing season</td>
<td>Target new acquisitions from undersampled times</td>
</tr>
<tr>
<td>Body size</td>
<td>Small / uncharismatic taxa underrepresented; larger / charismatic taxa overrepresented</td>
<td>Ensure new acquisitions are made using standard ecological sampling protocols.</td>
</tr>
<tr>
<td>Abundance</td>
<td>Only the most common species are likely to have museum lot frequency reflect ecological frequency and abundance</td>
<td>Ensure new acquisitions are made using standard ecological sampling protocols.</td>
</tr>
<tr>
<td><strong>Labeling Errors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Misidentification</td>
<td>Present in ~20% of lots; most common in intermediate size classes and high-richness families. Flattens body size distribution and confuses conservation priorities</td>
<td>Expert on-site verification before digitized data released to public</td>
</tr>
<tr>
<td>Mixed lots</td>
<td>Present in ~4% of lots; most common in small organisms, diverse families and some collections. Leads to richness underreporting. Body size distribution unaffected</td>
<td>Expert on-site verification before digitized data released to public</td>
</tr>
<tr>
<td>Incorrect locality information</td>
<td>Generation of spurious records; limits conservation planning and taxonomic review</td>
<td>Error correction options range from historical research &amp; computer code to impossible</td>
</tr>
<tr>
<td>Incomplete locality information</td>
<td>Often cannot be fixed</td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----------------------</td>
<td></td>
</tr>
<tr>
<td>Most frequent in old lots, rare habitats / species, and where data is held in a separate location (e.g., collection notebook). Limits conservation planning and taxonomic review.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Misspelled information</td>
<td>Increases noise</td>
<td>Trained data entry staff / Computer code</td>
</tr>
<tr>
<td>Curation Errors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixed labels</td>
<td>Generation of spurious records and biogeographic / ecological noise</td>
<td>Expert on-site verification; may often be uncorrectable</td>
</tr>
<tr>
<td>Specimen degradation</td>
<td>Cannot accurately reverify old lots</td>
<td>Correct archival conditions</td>
</tr>
<tr>
<td>Label degradation / illegibility</td>
<td>Cannot accurately deduce old collection information</td>
<td>Correct archival conditions</td>
</tr>
</tbody>
</table>
Figure 1: Number of UTEP (left) and Perot (right) lots per county

Figure 2: Impact of log(county population), log(land area), and presence of national parks on number of museum lots. Dot diameter is linearly scaled to the number of lots/county, ranging from the biggest (398; El Paso) to least (0; 74 counties). Significance for the multiple linear regression is $p<0.0000000001$, $r^2=0.2858$. The significance for each independent variable is: log(1980 population): $p=0.00000012$; log(land area): $p=0.008148$; presence of national park: $p<0.00000001$. 
Figure 3: Unexplained variance vs. distance to repository from multiple linear regression using log(1980 population), log(land area) and presence of a national park as predictor variables.

Figure 4: Comparison of museum lots vs. ecological lots / individuals for log binned size classes (left) and cumulative frequencies across the body size spectrum (right). Tests for independence are based on log-likelihood ratio (left) and Kolmogorov–Smirnov test (right).
Figure 5: Spearman-rank correlation $p$-values for museum lot frequency vs. ecological sample frequency (left) and abundance (right) for all 98 jointly-held species (left-most dot) and after removal of the 1$^{st}$, 1$^{st}$-2$^{nd}$, 1$^{st}$-3$^{rd}$ .... 1$^{st}$-(n-2)$^{th}$ most abundant species. The rank order axis displays the cutoff between removed and retained species.

Figure 6: Cumulative Rank Frequency Distribution in log-log space for the number of misidentified lots per species across both museums.
Figure 7: Cumulative Rank Frequency Distributions in log-log space for the number of mixed lots species (left) and the number of additional records (right) per species across both museums.