Received: 8 February 2019 | Revised: 12 July 2019 | Accepted: 19 July 2019

DOI: 10.1111/geb.12995

ECOLOGICAL SOUNDING

Global Ecology and Biogeography A Journal of Macroecology

WILEY

Caveat consumptor notitia museo: Let the museum data user beware

Jeffrey C. Nekola[1] | Benjamin T. Hutchins[2] | Alison Schofield[3] | Briante Najev[3] | Kathryn E. Perez[3]

[1]Department of Botany and Zoology, Masaryk University, Brno, Czech Republic

[2]Edwards Aquifer Research & Data Center & Department of Biology, Texas State University, San Marcos, Texas

[3]Department of Biology, University of Texas Rio Grande Valley, Edinburg, Texas

Correspondence
Jeffrey C. Nekola, Department of Botany and Zoology, Masaryk University, Kotlářská 2, CZ-611 37 Brno, Czech Republic.
Email: nekola@sci.muni.cz

Funding information
Texas Parks and Wildlife Department, Grant/Award Number: TX T-155-R-1, F16AF01265; U.S. Fish and Wildlife Service

Editor: Shai Meiri

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 favour 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 upon 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

body size, data mining, misidentification, museum records, sampling bias, specimen labelling error

1 | 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, Kitching, Livermore, Simonsen, & Smith, 2012). Museums are actively promoting their use (Funk, 2018; Krishtalka & Humphrey, 2000) as an important resource to test a wide range of biodiversity hypotheses

(Graham, Ferrier, Huettman, Moritz, & Townsend-Peterson, 2004; Lavoie, 2013; Suarez & Tsutsui, 2004). The macroecological community has responded by beginning to mine and utilize such information to test hypotheses (e.g., Babin-Fenske, Anand, & Alarie, 2008; Economo, Narula, Friedman, Weiser, & Guénard, 2018; Foley et al., 2008; Ramirez-Villegas, Jarvis, & Touval, 2012).

It is unquestioned that museum collections represent an invaluable data source for research into systematics (Wen, Ickert-Bond, Appelhans, Dorr, & Funk, 2015), evolution (Àvila-Arcos et al., 2012), biological conservation (Drew, 2011) and environmental change (Pyke & Ehrlich, 2010). The idiosyncratic nature of museum accessions (Ponder, Carter, Flemons, & Chapman, 2001), however, raises a larger question about their ability to accurately reflect ecological and biogeographic patterns. Does bias and error in these data fundamentally limit their utility for drawing accurate conclusions? Or are they simply a nuisance that 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 naïve 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 analyse 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.

2 | POTENTIAL SCOPE OF THE PROBLEM

Minimally, three steps lie between an ecological pattern in the field and its representation as museum record data: collection, labelling and curation. Each is associated with potential biases or errors (Newbold, 2010; Soberón, Llorente, & Oñate, 2000):

2.1 | 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 (*Plathanthera praeclara* 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–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 under-represented. 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 a 1,500-km extent (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, Downey, & Nekola, 2007).

The taxonomic focus of many collectors also means that their collections often emphasize favourite 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 fieldwork 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 occur 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.

2.2 | Labelling 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 that 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 that 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* Delboutteville was reported from Roaring Springs, "Travies Country", Texas (Deboutteville, Coineau, & Serban, 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 south-eastern 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, August 23, 2019). 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).

2.3 | Curation errors

A final suite of potential problems is associated with activities following accession. While most relate to specimen or data loss through incorrect labelling 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 93,636 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.

3 | 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 t o 16% of sample localities 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 patterns. 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, Fullington, & Pratt, 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, Bieler, Shea, & Rosenberg, 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 onsite 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.

3.1 | Datasets

3.1.1 | Museum record data

We entered label data and verified identifications for all Texas-sourced lots, representing 2,582 from the Perot and 1,190 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 (Supporting Information S1).

3.1.2 | 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, 1,048 lots and 52,065 individuals (Supporting Information S2). 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 100–1,000 m^2 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 (Cameron & Pokryszko, 2005; Oggier, Zschokke, & Baur, 1998). Litter sampling was carried out 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, Pearce, & Randalana, 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 *c.* 500 ml of retained litter per site was slowly and completely dried in the laboratory and then passed through a standard sieve series [American Standard Test Mesh 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.

3.1.3 | 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 and Rocroi (2005) and Schileyko (2006). County-scale demographic and geographic data were obtained via the Texas Open Data Portal (https://data.texas.gov/).

3.2 | Collection biases

3.2.1 | 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 ($p < .00815$) predicted lot numbers per county and in total explained almost 30% of observed variance (Figure 2).

Because the two repositories are separated by 1,000 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 they provided the best linear model fit.

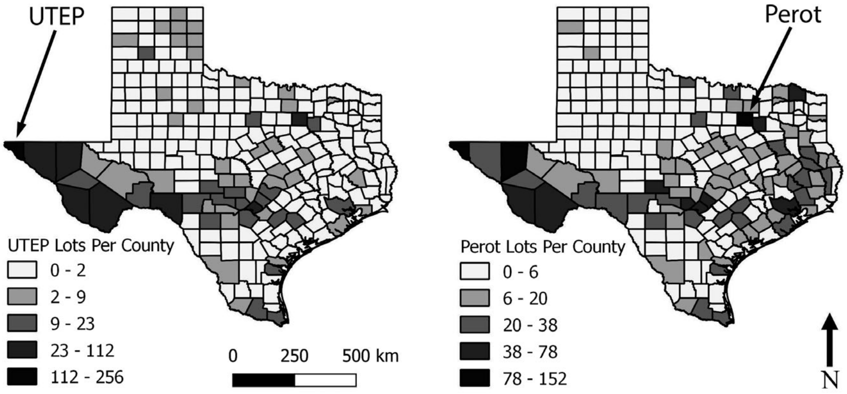


FIGURE 1 Number of University of Texas El Paso (UTEP; left) and Perot (right) lots per county

While distance to UTEP was a highly significant additional predictor ($p < .0000001$), explaining over 25% of remaining variance (Figure 3), distance to Perot Museum was not ($p = .624$). 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 ($p = .000723$), 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.

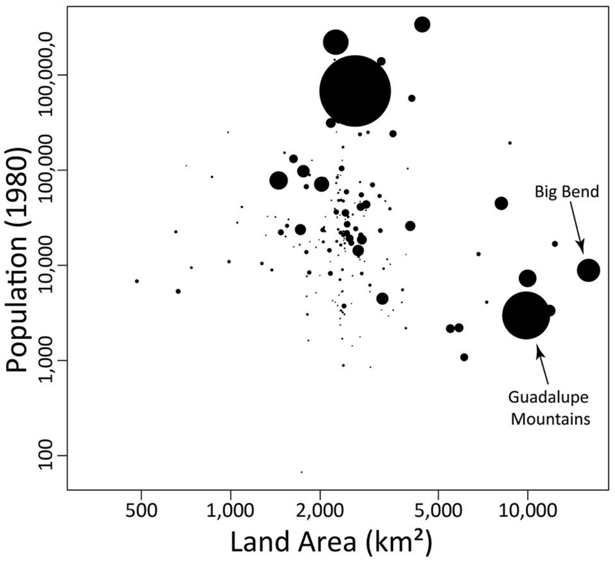


FIGURE 2 Impact of log(county population), log(land area) and presence of national parks on number of museum lots. Dot diameter is linearly scaled to the number of lots/county, ranging from the biggest (398; El Paso) to smallest (0; 74 counties). Significance for the multiple linear regression is $p < .0000000001$, $r^2 = .2858$. The significance for each independent variable is: log(1980 population): $p = .00000012$; log(land area): $p = .008148$; presence of national park: $p < .00000001$

3.2.2 | Collection date

We could not empirically test for this bias because: (a) only 950 records (*c.* 25%) provided collection date beyond year; (b) 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 were both <= 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 ($p = .003$; 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.

3.2.3 | 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 ($p < .000000001$) and Kolmogorov–Smirnov ($p < .000009$) tests demonstrated (Figure 4) that museum lots significantly under-represented the frequency of small and over-represented the frequency of large species as compared to both occurrences and total individuals encountered during ecological sampling.

3.2.4 | Relative species abundance

This potential bias was investigated through Spearman rank correlation on the total number of museum lots for each species

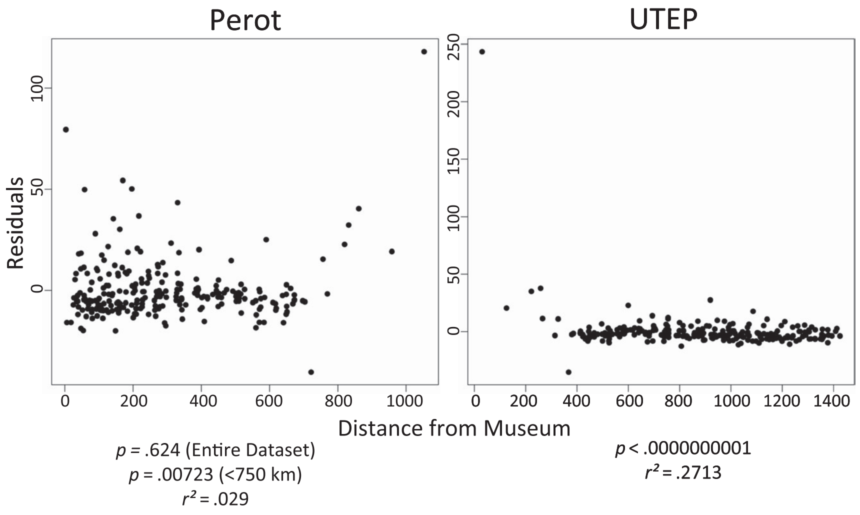
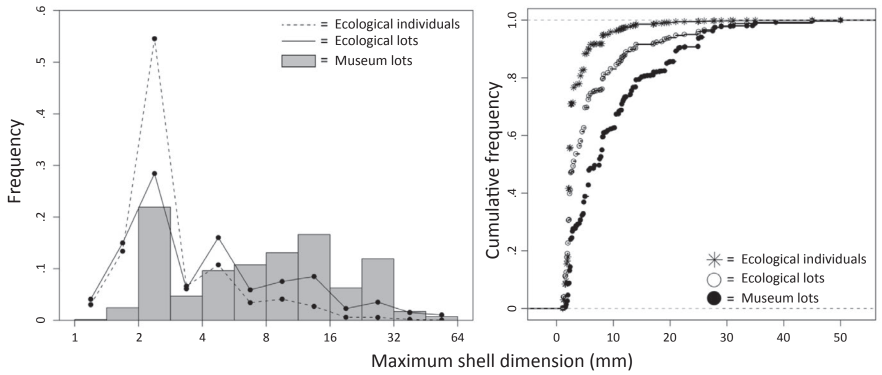


FIGURE 3 Unexplained variance versus distance to repository from multiple linear regression using log(1980 population), log(land area) and presence of a national park as predictor variables

FIGURE 4 Comparison of museum lots versus 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)



Museum lots versus Ecological lots:
$p < .000000001$
Museum lots versus Ecological individuals
$p < .000000001$

Museum lots versus Ecological lots:
$p = .0000087$
Museum lots versus Ecological individuals
$p < .000000001$

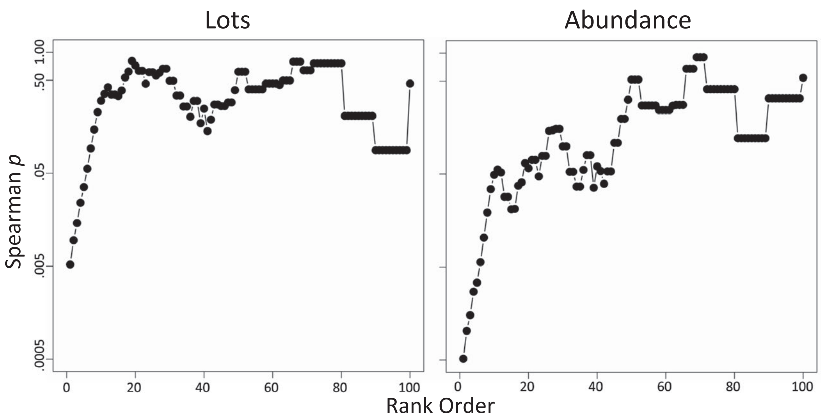
versus total number of lots and individuals encountered through ecological sampling. Only the 98 species present in both datasets were used. While the overall correlation was significant and positive (Figure 5) for museum lots versus ecological occurrences (*p* = .0052; rho = .28) and versus ecological abundances (*p* = .00052; rho = .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 that included the five (versus ecological lots) or eighteen (versus 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.

3.3 | Labelling errors

Given the effort that has been made to identify and correct geolocation/place name errors in museum record data (e.g., Guralnick, Hill, & Lane, 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 principal 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 uppermost left point represents the 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

FIGURE 5 Spearman rank correlation *p*-values for museum lot frequency versus ecological sample frequency (left) and abundance (right) for all 98 jointly held species (left-most dot) and after removal of the 1st, 1st–2nd, 1st–3rd.... 1st–(*n*–2)th most abundant species. The rank order axis displays the cut-off between removed and retained species



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 versus 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 versus verified identifications and between principal and additional shells in mixed lots was also documented through use of log-likelihood ratio contingency table tests.

3.3.1 | Lot misidentification

Over 20% of lots (769 out of the 3,772) 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 and 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 [*Glyphyalinia umbilicata* (Singley in Cockerell), *Hawaiia miniscula* (A. Binney), *Linisa texasiana* (Moricand), *Gastrocopta sterkiana* Pilsbry, *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 = .124) in misidentification rate between the two repositories (Table 1A). However, there was strong dependence (p < .00000001) in body size with highest misidentification rates (21–27%) occurring in the three intermediate size classes (Table 1B). Strong dependence (p < .000000001) 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 < .000000001) more uniform shell size distribution (Table 1D).

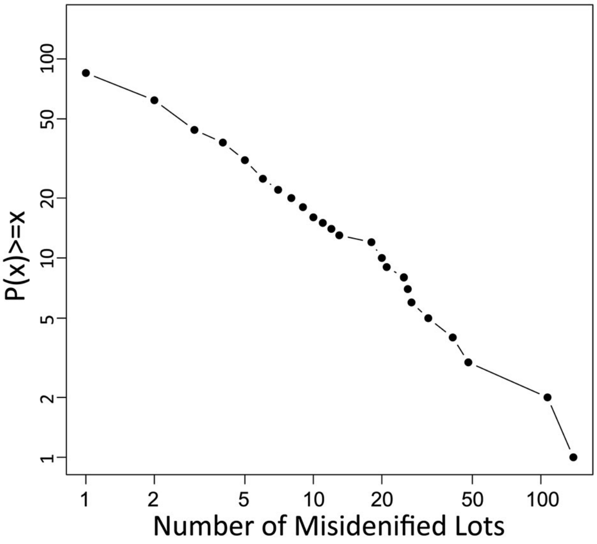


FIGURE 6 Cumulative rank frequency distribution in log–log space for the number of misidentified lots per species across both museums

3.3.2 | Lot mixing

Almost 4% of lots (149 out of the 3,772) possessed multiple species, with a total of 196 additional occurrences being recorded. Again, this led to some critical errors with two species – *Gastrocopta rogersensis* Nekola & Coles and *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 [*Hawaiia miniscula*, *Gastrocopta contracta* (Say), *Gastrocopta procera* (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 [*Gastrocopta sterkiana*, *Gastrocopta pentodon*, *Hawaiia miniscula*, *Gastrocopta pellucida*, *Helicodiscus singleyanus* (Pilsbry)] 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 versus UTEP collection (p = .00000053; Table 2A), and was also greatly influenced (p < .00000001) by shell size with the lowest rates (*c.* 1%) occurring at the largest and highest rates occurring (9%) in the smallest size class (Table 2B). Strong dependence (p < .000000001) 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 = .50) in shell size distribution of principal versus additional shells in mixed lots (Table 2D).

3.3.3 | Erroneous and missing location data

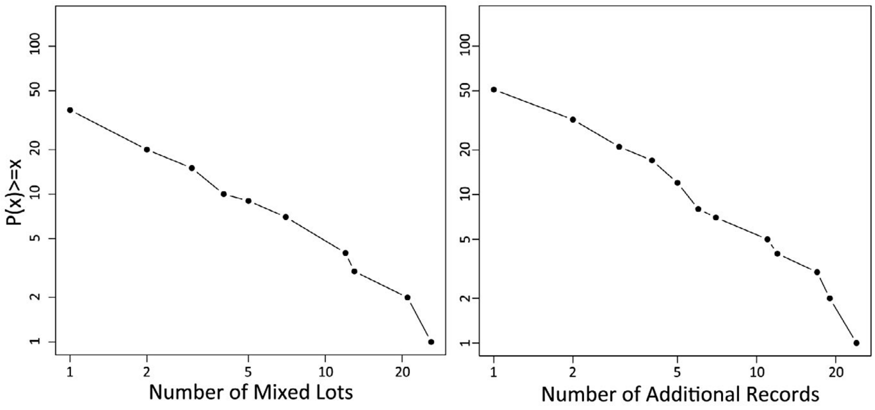
While we chose to not explicitly investigate rates of misspelled/erroneous geographic place names or incorrect georeferencing we note that these types of errors were common. Also of concern were instances where label information was incorrectly transcribed from collection notebooks. Perhaps the most egregious example was two dozen UTEP Hutchinson County lots collected on 19 June 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 11 species either unknown from

TABLE 1 Factors associated with misidentification

A. Repository								
Repository	**Perot**				**UTEP**			
Correct	2,038				965			
Misidentified	544				225			
Percent	21.1				18.9			
B. Maximum shell dimension								
Maximum dimension (mm)	< 2.5	2.5–< 5		5–< 10	10–< 20		20+	
Correct	488	619		745	694		457	
Misidentified	57	229		211	184		88	
Percent wrong	10.5	27.0		22.1	21.0		16.2	
C. Family richness within the North American fauna								
Number of species	1	2–3	4–7	8–15	16–31	32–63	64–127	128–256
Correct	3	19	183	575	917	264	195	847
Misidentified	0	7	8	70	257	149	110	168
Percent wrong	0.0	26.9	4.2	10.9	21.9	36.1	36.1	16.6
D. Impact on body size distribution								
Maximum dimension (mm)	< 2.5	2.5–< 5		5–< 10	10–< 20		20+	
Original ID	141	137		193	197		79	
Verified ID	53	221		206	175		92	

Note: UTEP = University of Texas El Paso.(A. Repository) Log-likelihood test statistic = 2.369; *df* = 1; *p*-value = .1238007. (B. Maximum shell dimension) Log-likelihood test statistic = 68.1257; *df* = 4; *p*-value << .000000001. (C. Family richness within the North American fauna) Log-likelihood test statistic = 188.13; *df* = 7; *p*-value << .000000001. (D. Impact on body size distribution) Log-likelihood test statistic = 64.0225; *df* = 4; *p*-value << .000000001.

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



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.

4 | 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 favour 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 favoured. 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 patterns. 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 4% 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 under-reporting of species richness from site to regional scales. In addition, we saw label information that 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 that 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, Peterson, Wisniewski, & Johnson, 2011), 20% for bird holdings in the Šariš Museum (Mikula, Csanády, & Hromada, 2018), 20% for fungi used in molecular analyses (Bridge, Roberts, Spooner, & Panchal, 2003) and 10% for germplasm holdings of *Citrullus* (watermelons) in international seed banks (Guzzon & Ardenghi, 2018). 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, Harris, Filer, Wood, & Scotland, 2015). Such misidentification errors have led to profound confusion in phylogenetic tree interpretation (Ó Foighil, Lee, Campbell, & Clark, 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

TABLE 2 Factors associated with lot mixing

A. Repository								
Repository	**Perot**				**UTEP**			
Unmixed	2,454				1,169			
Mixed	128				21			
Percent	5.0				1.8			
B. Maximum shell dimension								
Maximum dimension (mm)	< 2.5	2.5–< 5		5–< 10	10–< 20		20+	
Unmixed	495	791		928	869		540	
Mixed	50	57		28	9		5	
Percent mixed	9.2	6.7		2.9	1.0		0.9	
C. Family richness within the North American fauna								
Number of species	1	2–3	4–7	8–15	16–31	32–63	64–127	128–256
Unmixed	3	26	188	639	1,099	396	279	993
Mixed	0	0	3	6	75	17	26	22
Percent mixed	0.0	0.0	1.6	0.9	6.4	4.1	8.5	2.3
D. Body size distribution of target versus mixed species within a lot								
Maximum dimension (mm)	< 2.5	2.5–< 5		5–< 10	10–< 20		20+	
Target	50	57		28	9		5	
Mixed	76	71		38	9		2	

Note: UTEP = University of Texas El Paso.(A. Repository) Log-likelihood test statistic = 25.1626; *df* = 1; *p*-value = .00000053. (B. Maximum shell dimension) Log-likelihood test statistic = 92.889; *df* = 4; *p*-value << .000000001. (C. Family richness within the North American fauna) Log-likelihood test statistic = 63.300; *df* = 7; *p*-value << .000000001. (D. Body size distribution of target versus mixed species within a lot) Log-likelihood test statistic = 3.36457; *df* = 4; *p*-value = .4987669.

TABLE 3 Summary of museum data issues and potential remediations

Type	Effect	Potential remediation
Collection biases		
Collection location	Collections more likely from counties with greater human population, more land area, iconic destinations, and nearer to repository	Target new acquisitions from undersampled areas
Collection date	Uneven sampling across year/ growing season	Target new acquisitions from undersampled times
Body size	Small/ uncharismatic taxa under-represented; larger/ charismatic taxa over-represented	Ensure new acquisitions are made using standard ecological sampling protocols.
Abundance	Only the most common species are likely to have museum lot frequencies that reflect ecological frequency and abundance	Ensure new acquisitions are made using standard ecological sampling protocols.
Labelling errors		
Misidentification	Present in *c.* 20% of lots; most common in intermediate size classes and high-richness families. Flattens body size distribution and confuses conservation priorities	Expert onsite verification before digitized data released to public
Mixed lots	Present in *c.* 4% of lots; most common in small organisms, diverse families and some collections. Leads to richness underreporting. Body size distribution unaffected	Expert onsite verification before digitized data released to public
Incorrect locality information	Generation of spurious records; limits conservation planning and taxonomic review	Error correction options range from historical research and computer code to impossible
Incomplete locality information	Most frequent in old lots, rare habitats/species, and where data are held in a separate location (e.g., collection notebook). Limits conservation planning and taxonomic review	Often cannot be fixed
Misspelled information	Increases noise	Trained data entry staff/computer code
Curation errors		
Mixed labels	Generation of spurious records and biogeographic/ecological noise	Expert onsite verification; may often be uncorrectable
Specimen degradation	Cannot accurately reverify old lots	Correct archival conditions
Label degradation/ illegibility	Cannot accurately deduce old collection information	Correct archival conditions

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 that provide accurate species-level diagnoses (Horsáková, Nekola, & Horsák, 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 favour of areas with denser human populations could well confound analyses and lead to incorrect interpretations.

5 | 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, Šizling, Boyer, & Storch, 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: (a) include undersampled regions, time periods, body size classes, taxa groups, etc., and (b) 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 typographic errors, outdated taxonomy, geographic 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 onsite 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, Haapala, Mononen, Pajari, & Saarenmaa, 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 and 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, M'Gonigle, & Resh, 2014), and accurate estimates of environmental range (Phillips, Anderson, & Schapire, 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 re-evaluated. 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, Tarnowsky, Bevans, Kirchgessnern, & Thiers, 2012).

ACKNOWLEDGMENTS

We thank the Perot and UTEP museums for access to specimens, particularly staff members Karen Morton, Chelsea Herrod and Teresa J. Mayfield. This project was supported by Texas Parks and Wildlife Department through the State Wildlife Grant Program, funded by the U.S. Fish and Wildlife Service (Grant #: TX T-155-R-1, F16AF01265) and the College of Sciences at The University of Texas Rio Grande Valley. Parwinder Grewal, Bill Elliott, Lillian E. Perez, Frederic Zaidan, Armando Salinas and Juan G. Perez provided essential logistical support for this project.

DATA ACCESSIBILITY

Because analysed data include specific locality information for many protected species, and because access to private land was often granted only after agreeing to not publicly 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.

ORCID

Jeffrey C. Nekola https://orcid.org/0000-0001-6073-0222

REFERENCES

Ávila-Arcos, M. C., Ho, S. Y., Ishida, Y., Nikolaidis, N., Tsangaras, K., Hönig, K., … Willerslev, E. (2012). One hundred twenty years of koala retrovirus evolution determined from museum skins. *Molecular Biology and Evolution*, *30*, 299–304.

Babin-Fenske, J., Anand, M., & Alarie, Y. (2008). Rapid morphological change in stream beetle museum specimens correlates with climate change. *Ecological Entomology*, *33*, 646–651.

Baker, F. C. (1939). *Fieldbook of Illinois land snails*. Urbana: Illinois Natural History Survey Division, Illinois Department of Registration and Education.

Ball-Damerow, J. E., M'Gonigle, L. K., & Resh, V. H. (2014). Changes in occurrence, richness, and biological traits of dragonflies and damselflies (Odonata) in California and Nevada over the past century. *Biodiversity and Conservation*, *23*, 2107–2126.

Barker, G. M. (2001). Gastropods on land: Phylogeny, diversity, and adaptive morphology. In G. M. Barker (Ed.), *The biology of terrestrial molluscs* (pp. 1–146). New York: CABI.

Bertone, M. A., Blinn, R. L., Stanfield, T. M., Dew, K. J., Seltmann, K. C., & Deans, A. R. (2012). Results and insights from the NCSU insect museum GigaPan project. *ZooKeys*, *209*, 115–132.

Blagoderov, V., Kitching, I., Livermore, L., Simonsen, T., & Smith, V. (2012). No specimen left behind: Industrial scale digitization of natural history collections. *ZooKeys*, *209*, 133–146.

Bouchet, P., & Rocroi, J.-P. (2005). Classification and nomenclator of gastropod families. *Malacologia*, *47*, 1–397.

Bridge, P. D., Roberts, P. J., Spooner, B. M., & Panchal, G. (2003). On the unreliability of published DNA sequences. *New Phytologist*, *160*, 43–48.

Cameron, R. A. D., & Pokryszko, B. M. (2005). Estimating the species richness and composition of land mollusc communities. *Journal of Conchology*, *38*, 529–547.

Cheatum, E. P., Fullington, R. W., & Pratt, W. L. (1974). *The aquatic and land Mollusca of Texas. Part 3*. Dallas, TX: Dallas Museum of Natural History.

Crovello, T. J. (1967). Problems in the use of electronic data processing in biological collections. *Taxon*, *16*, 481–494.

Deboutteville, C. D., Coineau, N., & Serban, E. (1975). Découverte de la famille des Parabathynellidae (Bathynellacea) en Amérique du Nord: *Texanobathynella bowmani* ngn. sp. *Comptes Rendus De L'académie Des Sciences*, *Série D*, *280*, 2223–2226.

Drew, J. (2011). The role of natural history institutions and bioinformatics in conservation biology. *Conservation Biology*, *25*, 1250–1252.

Economo, E. P., Narula, N., Friedman, N. R., Weiser, M. D., & Guénard, B. (2018). Macroecology and macroevolution of the latitudinal diversity gradient in ants. *Nature Communications*, *9*, 1778.

Emberton, K. C., Pearce, T. A., & Randalana, R. (1996). Quantitatively sampling land-snail species richness in Madagascan rainforests. *Malacologia*, *38*, 203–212.

Foley, D. H., Weitzman, A. L., Miller, S. E., Faran, M. E., Rueda, L. M., & Wilkerson, R. C. (2008). The value of georeferenced collection records for predicting patterns of mosquito species richness and endemism in the Neotropics. *Ecological Entomology*, *33*, 12–23.

Funk, V. A. (2018). Collections-based science in the 21st Century. *Journal of Systematics and Evolution*, *56*, 175–193.

Goodwin, Z. A., Harris, D. J., Filer, D., Wood, J. R. I., & Scotland, R. W. (2015). Widespread mistaken identity in tropical plant collections. *Current Biology*, *25*, R1057–R1069.

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C., & Townsend-Peterson, A. (2004). New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution*, *19*, 497–503.

Guralnick, R. P., Hill, A. W., & Lane, M. (2007). Towards a collaborative, global infrastructure for biodiversity assessment. *Ecology Letters*, *10*, 663–672.

Guzzon, F., & Ardenghi, N. M. G. (2018). Could taxonomic misnaming threaten the ex situ conservation and usage of plant genetic resources? *Biodiversity and Conservation*, *27*, 1157–1172.

Hanna, G. D., & Smith, A. G. (1954). Rediscovery of two Californian land snails. *The Nautilus*, *67*, 69–76.

Horsáková, V., Nekola, J. C., & Horsák, M. (2019). When is a "cryptic" species not a cryptic species: A consideration from the Holarctic micro-landsnail genus *Euconulus* (Gastropoda: Stylommatophora). *Molecular Phylogenetics and Evolution*, *132*, 307–320.

Krishtalka, K., & Humphrey, P. S. (2000). Can natural history museums capture the future? *BioScience*, *50*, 611–617.

Lavoie, C. (2013). Biological collections in an ever changing world: Herbaria as tools for biogeographical and environmental studies. *Perspectives in Plant Ecology, Evolution and Systematics*, *15*, 68–76.

McClain, C. (2011). The mass extinction of scientists who study species. *Wired Science*, *19/01*. Retrieved from https://www.wired.com/2011/01/extinction-of-taxonomists/

Meier, R., & Dikow, T. (2004). Significance of specimen databases from taxonomic revisions for estimating and mapping the global species diversity of invertebrates and repatriating reliable specimen data. *Conservation Biology*, *18*, 478–488.

Mikula, P., Csanády, A., & Hromada, M. (2018). A critical evaluation of the exotic bird collection of the Šariš Museum in Bardejov, Slovakia. *ZooKeys*, *776*, 139–152.

Neck, R. W. (1980). Habitat notes on *Gastrocopta riograndensis* Sterki. *The Veliger*, *23*, 180–182.

Nekola, J. C. (2009). *Conservation prioritization of the Ontario and Quebec land snail faunas* (Final Report). Ottawa: Committee on the Status of Endangered Wildlife in Canada.

Nekola, J. C. (2014). Overview of the North American terrestrial gastropod fauna. *American Malacological Bulletin*, *32*, 225–235.

Nekola, J. C., & Coles, B. F. (2010). Pupillid land snails of eastern North America. *American Malacological Bulletin*, *28*, 29–57.

Nekola, J. C., Šizling, A. L., Boyer, A. G., & Storch, D. (2008). Artifactions in the log-transformation of species abundance distributions. *Folia Geobotanica*, *43*, 259–468.

Newbold, T. (2010). Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography*, *34*, 3–22.

Newman, M. E. J. (2005). Power laws, Pareto distributions, and Zipf's law. *Contemporary Physics*, *46*, 323–351.

Ó Foighil, D., Lee, T., Campbell, D. C., & Clark, S. A. (2009). All voucher specimens are not created equal: A cautionary tale involving North American pleurocerid gastropods. *Journal of Molluscan Studies*, *75*, 305–306.

Oggier, P., Zschokke, S., & Baur, B. (1998). A comparison of three methods for assessing the gastropod community in dry grasslands. *Pedobiologia*, *42*, 348–357.

Palmer, M. W. (1995). How should one count species? *Natural Areas Journal*, *15*, 124–135.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*, 231–259.

Pilsbry, H. A. (1948). *Land Mollusca of North America (North of Mexico)*. Monograph #3. PA: Academy of Natural Sciences of Philadelphia.

Ponder, W. F., Carter, G. A., Flemons, P., & Chapman, R. R. (2001). Evaluation of museum collection data for use in biodiversity assessment. *Conservation Biology*, *15*, 648–657.

Preston, F. W. (1948). The commonness, and rarity, of species. *Ecology*, *29*, 254–283.

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.

Ramirez-Villegas, J., Jarvis, A., & Touval, J. (2012). Analysis of threats to South American flora and its implications for conservation. *Journal for Nature Conservation*, *20*, 337–348.

Schileyko, A. A. (2006). Treatise on recent terrestrial pulmonate molluscs, Part 13. *Ruthenica*, Supplement *2*, 1765–1906.

Schlicht, D. W., Downey, J. C., & Nekola, J. C. (2007). *The butterflies of Iowa*. Iowa City, IAUniversity of Iowa Press.

Shea, C. P., Peterson, J. T., Wisniewski, J., & Johnson, N. (2011). Misidentification of freshwater mussel species (Bivalvia:Unionidae): Contributing factors, management implications, and potential solutions. *Journal of the North American Benthological Society*, *30*, 446–458.

Sierwald, P., Bieler, R., Shea, E. K., & Rosenberg, G. (2018). Mobilizing mollusks: Status update on mollusk collections in the USA and Canada. *American Malacological Bulletin*, *36*, 177–214.

Slotten, R. A. (2004). *The heretic in Darwin's court: The life of Alfred Russel Wallace*. New York, NY: Columbia University Press.

Smith, A. G. (1957). Snails from California caves. *Proceedings of the California Academy of Sciences*, *29*, 21–46.

Soberón, J. M., Llorente, J. B., & Oñate, L. (2000). The use of specimen-label databases for conservation purposes: An example using Mexican Papilionid and Pierid butterflies. *Biodiversity and Conservation*, *9*, 1441–1466.

Sokal, R. R., & Rohlf, F. J. (1981). *Biometry: The principles and practice of statistics in biological research* (2nd ed.). New York, NY: Freeman.

Suarez, A. V., & Tsutsui, N. D. (2004). The value of museum collections for research and society. *BioScience*, *54*, 66–74.

Tegelberg, R., Haapala, J., Mononen, T., Pajari, M., & Saarenmaa, H. (2012). The development of a digitising service centre for natural history collections. *ZooKeys*, *209*, 75–86.

Tennent, N. H., & Baird, T. (1985). The deterioration of Mollusca collections: Identification of shell efflorescence. *Studies in Conservation*, *30*, 73–85.

Tulig, M., Tarnowsky, N., Bevans, M., Kirchgessnern, A., & Thiers, B. M. (2012). Increasing the efficiency of digitization workflows for herbarium specimens. *ZooKeys*, *209*, 103–113.

Vilgalys, R. (2003). Taxonomic misidentification in public DNA databases. *New Phytologist*, *160*, 4–5.

Wen, J., Ickert-Bond, S. M., Appelhans, M. S., Dorr, L. J., & Funk, V. A. (2015). Collections - based systematics: Opportunities and outlook for 2050. *Journal of Systematics and Evolution*, *53*, 477–488.

BIOSKETCH

Jeffrey C. Nekola is an associate professor in the Department of Botany and Zoology at Masaryk University. His works span the range of biodiversity studies, ranging from molecular genetics to macroecology and general theory.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Nekola JC, Hutchins BT, Schofield A, Najev B, Perez KE. *Caveat consumptor notitia museo*: Let the museum data user beware. *Global Ecol Biogeogr.* 2019;28:1722–1734. https://doi.org/10.1111/geb.12995