Assessment of North American arthropod collections: Prospects and challenges for addressing biodiversity research (#39532)

First submission

Guidance from your Editor

Please submit by **16 Aug 2019** for the benefit of the authors (and your $200 publishing discount).

Structure and Criteria
Please read the 'Structure and Criteria' page for general guidance.

Custom checks
Make sure you include the custom checks shown below, in your review.

Author notes
Have you read the author notes on the guidance page?

Raw data check
Review the raw data. Download from the location described by the author.

Image check
Check that figures and images have not been inappropriately manipulated.

Privacy reminder: If uploading an annotated PDF, remove identifiable information to remain anonymous.

Files
Download and review all files from the materials page.

10 Figure file(s)
4 Table file(s)

Custom checks

Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- Original primary research within Scope of the journal.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- Impact and novelty not assessed. Negative/inconclusive results accepted. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- All underlying data have been provided; they are robust, statistically sound, & controlled.
- Speculation is welcome, but should be identified as such.
- Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

Assessment of North American arthropod collections: Prospects and challenges for addressing biodiversity research

Neil S Cobb [Corresp., 1], **Lawrence F Gall** [2], **Jennifer M Zaspel** [3], **Nicolas J Dowdy** [3, 4], **Lindsie M McCabe** [5], **Akito Y Kawahara** [6]

[1] Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ, United States

[2] Entomology Division, Yale Peabody Museum of Natural History, Yale University, New Haven, Connecticut, United States

[3] Department of Zoology, Milwaukee Public Museum, Milwaukee, Wisconsin, United States

[4] Department of Biology, Wake Forest University, Winston-Salem, North Caroloina, United States

[5] Department of Biological Sciences, Northern Arizona University, Flagstaff, Arizona, United States

[6] Florida Museum of Natural History, ,, University of Florida, Gainesville, Florida, United States

Corresponding Author: Neil S Cobb
Email address: neil.cobb@nau.edu

Over 300 million arthropod specimens are housed in North American natural history collections. These collections represent a "last hidden treasure trove" of biodiversity – 95% of the specimen label data have yet to be transcribed for research, and less than 2% of the specimens have been imaged. Specimen labels contain crucial information to determine species distributions over time and are essential for understanding patterns of ecology and evolution, which will help assess the growing biodiversity crisis driven by global change impacts. Specimen images offer indispensable insight and data for analyses of traits, and ecological and phylogenetic patterns of biodiversity. Here, we review North American arthropod collections using two key metrics, specimen holdings and digitization efforts, to assess the potential for collections to provide needed biodiversity data. We include data from 223 arthropod collections in North America, with an emphasis on the United States. Our specific findings are as follows: 1. The majority of North American natural history collections (88%) and specimens (89%) are located in the United States. Canada has comparable holdings to the United States relative to its estimated biodiversity. Mexico has made the furthest progress in terms of digitization, but its specimen holdings should be increased to reflect the estimated higher Mexican arthropod diversity. The proportion of North American collections that has been digitized, and the number of digital records available per species, are both much lower for arthropods when compared to chordates and plants. The National Science Foundation's decade-long ADBC program (Advancing Digitization of Biological Collections) has been transformational in promoting arthropod digitization. However, even if this program became permanent, at current rates, by the year 2050 only 38% of the existing arthropod specimens would be digitized, and

less than 1% would have associated digital images. 3. The number of specimens in collections has increased by approximately 1% per year over the past 30 years. We propose that this rate of increase is insufficient to provide enough data to address biodiversity research needs, and that arthropod collections should aim to triple their rate of new specimen acquisition. 4. The collections we surveyed in the United States vary broadly in a number of indicators. Collectively, there is depth and breadth, with smaller collections providing regional depth and larger collections providing greater global coverage. 5. Increased coordination across museums is needed for digitization efforts to target taxa for research and conservation goals and address long-term data needs. Two key recommendations emerge, collections should significantly increase both their specimen holdings and their digitization efforts to empower continental and global biodiversity data pipelines, and stimulate downstream research.

Assessment of North American Arthropod Collections: Prospects and Challenges for Addressing Biodiversity Research

NEIL S. COBB [1] , LAWRENCE F. GALL [2], ,JENNIFER M. ZASPEL [3,4] ,NICOLAS J. DOWDY [3,5] , LINDSIE M. MCCABE [1], AKITO Y. KAWAHARA [6]

[1]Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ, USA
[2] Entomology Division, Yale Peabody Museum of Natural History, New Haven, CT, USA
[3] Department of Zoology, Milwaukee Public Museum, Milwaukee, WI, USA
[4] Department of Entomology, Purdue University, West Lafayette, IN, USA
[5] Department of Biology, Wake Forest University, Winston-Salem, NC, USA
[6] Florida Museum of Natural History, University of Florida, Gainesville, FL, USA

Corresponding Author:
Neil S. Cobb
617 S Beaver St, Flagstaff, Arizona, 86011, USA
Email address: Neil.Cobb@nau.edu

Abstract

Over 300 million arthropod specimens are housed in North American natural history collections. These collections represent a “vast hidden treasure trove” of biodiversity – 95% of the specimen label data have yet to be transcribed for research, and less than 2% of the specimens have been imaged. Specimen labels contain crucial information to determine species distributions over time and are essential for understanding patterns of ecology and evolution, which will help assess the growing biodiversity crisis driven by global change impacts. Specimen images offer indispensable insight and data for analyses of traits, and ecological and phylogenetic patterns of biodiversity. Here, we review North American arthropod collections using two key metrics, specimen holdings and digitization efforts, to assess the potential for collections to provide needed biodiversity data. We include data from 223 arthropod collections in North America, with an emphasis on the United States. Our specific findings are as follows:

1. The majority of North American natural history collections (88%) and specimens (89%) are located in the United States. Canada has comparable holdings to the United States relative to its estimated biodiversity. Mexico has made the furthest progress in terms of digitization, but its specimen holdings should be increased to reflect the estimated higher Mexican arthropod diversity. The proportion of North American collections that has been digitized, and the number of digital records available per species, are both much lower for arthropods when compared to chordates and plants.
2. The National Science Foundation's decade-long ADBC program (Advancing Digitization of Biological Collections) has been transformational in promoting arthropod digitization. However, even if this program became permanent, at current rates, by the year 2050 only 38% of the existing arthropod specimens would be digitized, and less than 1% would have associated digital images.
3. The number of specimens in collections has increased by approximately 1% per year over the past 30 years. We propose that this rate of increase is insufficient to provide enough

data to address biodiversity research needs, and that arthropod collections should aim to triple their rate of new specimen acquisition.
4. The collections we surveyed in the United States vary broadly in a number of indicators. Collectively, there is depth and breadth, with smaller collections providing regional depth and larger collections providing greater global coverage.
5. Increased coordination across museums is needed for digitization efforts to target taxa for research and conservation goals and address long-term data needs.

Two key recommendations emerge, collections should significantly increase both their specimen holdings and their digitization efforts to empower continental and global biodiversity data pipelines, and stimulate downstream research.

Introduction

Arthropod Natural History Collections

With more than one million described species, Arthropoda is the most taxonomically and ecologically diverse animal phylum, comprising over half of both North American and global animal species diversity. Arthropods include insects, arachnids, and crustaceans. Insects and arachnids are pervasive in non-marine environments, and crustaceans dominate most marine environments. Arthropods are fundamental to ecosystem function and impact humans both positively and negatively (McIntyre 2000). Arthropods are declining rapidly due to recent anthropogenic disturbance, such as climate change, noise and light pollution (Janzen & Hallwachs 2019; Lister & Garcia 2018; Sánchez-Bayo & Wyckhuys 2019), underscoring an urgency in documenting their life histories and geographic distributions and preserving specimens for future research.

Here we examine 223 collections of arthropods in North America (Canada, Mexico and United States, including territories) that vary in size, governance, and locality (Fig 1). Our overarching objectives include characterizing different types of arthropod collections, articulating challenges specific to arthropod collections, and assessing digitization efforts to date with a focus on meeting research data needs. We conducted analyses to examine broad scale trends concerning holdings and digitization efforts for all three countries but emphasize the United States (US) because we have the most complete data for that region. Collections assessed ranged from specialized small collections (~500 specimens) to the United States National Museum (USNM) collection with 35 million specimens. Most of the North American collections have dedicated websites and are housed in universities, public museums, and repositories for government programs.

Our focus is on arthropod collections, which have large samples of insects (i.e., 96% of arthropod records discussed herein are for insects). At least 40% of North American insect collections curate additional arthropod groups including Arachnida, Chilopoda, Crustacea, Diplopoda, and Entognatha(SCAN 2019). A number of collections curate invertebrates sensu lato, but we only surveyed those if they included insects. Additionally, we did not attempt to enumerate parasitic arthropods held in vertebrate collections (typically curated as data associated with vertebrate host specimens).

We also summarize digitization efforts among the four tiers defined by collection size. Most small entomology collections are located within college and university departments, where the person responsible is a faculty member with a variety of additional responsibilities. These

collections are often (1) focused on local fauna and/or reflect the particular interests of the curator(s), (2) managed and curated at their discretion, (3) lacking in dedicated institutional IT support, and (4) possibly supported by nominal budgets and/or students who receive credit for their participation. Larger entomology collections are usually housed in museums that are either free-standing institutions or institutions affiliated with a larger university. These collections are typically (1) of regional or worldwide scope, (2) managed by a dedicated curator and/or collection manager, (3) have access to institutional IT support, and (4) are supported by longer-term budget commitments and access to institutional personnel and related resources. Although the potential capacity to produce digital products at larger collections is much greater than at small collections, the former are also embedded within a broader administrative infrastructure which often present other challenges.

Defining Digitization for Arthropod Collections

Digitization is a term whose definition has been expanding in scope as technology allows more extraction of data from specimens (Nelson & Ellis 2018; Short et al. 2018; Watanabe 2019). We define digitization in the context of arthropod specimens as encompassing: (1) transcription of specimen labels into a database; (2) georeferencing localities (determining latitude/longitude); (3) capturing habitus image(s); and (4) vetting species-level identifications. These four elements of digitization are required to make records useful for most research purposes. Current digitization efforts focus almost exclusively on transcribing label data from specimens and georeferencing associated locality information (some efforts include capturing historical field notes e.g.,Nufio et al. (2010)). Most collections capture habitus images for exemplar specimens, but less than 1% of specimens have had a general habitus image recorded. Even fewer specimens have associated genetic data. There are some examples of collections linking genetic data to specimens (Short et al. 2018), or molecular tissue vouchers to specimens (Cho et al. 2016), but there is still rudimentary linkage between most genetic data in the Barcode of Life Datasystems (Ratnasingham & Hebert 2007) and in similar genomic repositories and specimen occurrence databases.

To achieve the highest value for scientific research, digitization should extract all possible information from specimens i.e., the "extended specimen" (Thiers et al. 2019) ~~sensu~~ including morphological, anatomical, molecular, and possibly even metabolomic data. As technology advances and becomes more accessible, our ability to obtain massive amounts of data from specimens will rapidly increase. For example, recent studies have captured phenotypic trait data from arthropod specimens to examine response to environmental change over time (Kharouba et al. 2018; McLean et al. 2016). Morphological traits in insects are also beginning to be assessed via automated workflows for 3D modelling derived from multi-angle imaging (Ströbel et al. 2018) as well as from microCT data (van de Kamp et al. 2015).

Importance of Specimen-based data for Biodiversity Research

In the past two decades digitized specimen records have become an invaluable resource for biodiversity and conservation research. Plant and vertebrate collections have spearheaded this effort (Bakker 2017; Bebber et al. 2010; Besnard et al. 2014; Bieker & Martin 2018; Braun & Wann 2017; Cook et al. 2014; Creley 2016; Davis et al. 2015; Greve et al. 2016; Guralnick & Constable 2010; Hart et al. 2014; Primack & Gallinat 2017; Schmitt et al. 2018; Willis et al.

2017). Other natural history collections have followed the lead of plants and vertebrates (Brooks et al. 2014; Lawson et al. 2018). Digitization is of benefit to collections by allowing them to share their holdings with larger audiences, and opening new avenues for large-scale research and public engagement (Ellwood et al. 2015; Ellwood et al. 2018; Nelson & Ellis 2018; Spear et al. 2017). Digitization also promotes collaborations among collections and integrated data at regional (Belitz et al. 2018; Sikes et al. 2016) and continental scales (Seltmann et al. 2017; Weirauch et al. 2017). Coordinated efforts to digitize arthropod collections across the US has resulted in an influx of specimen-level data and high-resolution images to online repositories (e.g., SCAN-BIF, iDigBio). This in turn offers great potential to address an array of environmental issues such as climate change, impacts of human land use, agricultural intensification and the spread of human and animal disease, and the role of arthropods in ecosystem services and crop/forest pest management (Belitz et al. 2018; Bell-Sakyi et al. 2018; Cook et al. 2014; Dunnum et al. 2017; Kharouba et al. 2018; Meineke et al. 2018). Specimen data are also emerging as critical pedagogical resources for science educators seeking to enhance teaching curricula and data literacy (Cook et al. 2014; Ellwood et al. 2019; Lacey et al. 2017; Monfils et al. 2017; Singer et al. 2018).

Recent reviews of arthropod natural history collections and emerging collections-based research have focused on different aspects of the importance of digitize specimens. Short et al. (2018) examined entomology collections in the “age of big data” with a focus on linking genetic data to specimens and technological advances in imaging. Bell-Sakyi et al. (2018) highlighted the importance and relevance of parasitic arthropod collections in understanding biotic interactions between disease vectors and their hosts. Kharouba et al. (2018) studied collections-based research addressing global change impacts, with examples relating to geographical distributions, phenology, phenotypic and genotypic traits. Other reviews have summarized the importance of collections in general, and raised concerns over their sustainability as fundamental providers of biodiversity data and the invaluable expertise of collection personnel, curators, and research associates for preparing data products to support convergent research (Krishtalka & Humphrey 2000; Thiers et al. 2019; Watanabe 2019).
For taxonomic groups other than arthropods that have been the focus of digitization efforts for some time, there are recent assessments of the efficacy of such efforts and the state of collections as it relates to producing relevant biodiversity data. For example, Singer et al. (2018) reviewed the major fish collections in the United States, updating holdings and digitization work over the last 22 years since the previous review by Poss & Collette (1995). Sierwald et al. (2018) provided a 40-year update on the survey of mollusk collections in the US and Canada since the previous review by Solem (1975). Our paper offers a comparable assessment of North American arthropod collections and establishes a baseline reference for future studies on museum collections.

Survey methodology

We began identifying collections and institutions for this survey in 2014 using the online resource "The Insect and Spider Collections of The World Website" (Evenhuis & Samuelson 2007). More than 90% of the institutions we surveyed acknowledged the presence of a collection on their website. For all collections, we used the estimate of holdings listed on the

collection website, in a few cases we followed up with direct correspondence to confirm holding size. We were reasonable confident that holding size did not include specimens in lots or large uncurated samples. Our list was compared periodically with several other resources: (1) a compendium of collections maintained by Song (2019); (2) collections listed in the database provided by the global registry of biodiversity repositories (Schindel & Cook 2018); and (3) collections that were established through the Symbiota Collection of Arthropods Network (SCAN) data portal at http://scan-bug.org (SCAN is a dedicated biodiversity portal that serves as an intermediate aggregator of data from 185 North American data providers) (SCAN 2019). Our final list included 223 collections from across North America.

For analysis of accumulated digital records, we restricted the survey to collections that have made their specimen data publicly available through SCAN, GBIF (https://www.gbif.org/) and/or iDigBio (Page et al. 2015). The SCAN data portal was queried on 22 October 2018 and on 24 January 2019, and results were cross-checked against both GBIF and iDigBio. The SCAN portal contained over 18 million records for North America during that three-month assessment period. We excluded 1.5 million records that represented observation-only or image-only records, and another 3 million records that had incomplete or unresolved taxonomic and/or locality data. This yielded a 13.4 million record sample, and we assumed error rates in species identifications and locality data did not differ appreciably among the collections that had contributed records. Data analyses were conducted using R scripts on a computing cluster at Northern Arizona University (http://nau.edu/hpc/).

For the United States collections, we placed each collection surveyed into one of four size classes that included all terrestrial and freshwater aquatic arthropod records. The four classes were: Tier 1 (< 100,000 specimens); Tier 2 (100,000 to < 1,000,000 specimens); Tier 3 (< 3,000,000 to 1,000,000 specimens); and Tier 4 (over 3,000,000 specimens). For temporal analysis, we defined a "historical record" as one where the collecting date was prior to 1965.

Results

Scope of North American Arthropod Collections and Digitization Efforts

Our survey of 223 arthropod collections from North America revealed that these collections currently house slightly more than 300 million specimens (Table S1), approximately triple the 93 million plant specimens estimated to be housed in North American herbaria (data from Index Herbariorum, March 2019, http://sweetgum.nybg.org/science/ih/). We were unable to determine an accurate estimate of the number of chordate (primarily vertebrates) specimens currently housed in North American collections, but that number is certainly smaller than for either plants or arthropods. These collection numbers do not strictly account for "specimen lots," where multiple individual specimens are collected and preserved together. This is routine practice for arthropods but less common for chordates and plants. Most of our data are for single dry-preserved specimens representing lots of n=1, and exclude immature arthropods, bulk samples, and other material typically stored in fluid or on slides as lots of n>1 (Sierwald et al. 2018). If we had been able to account for specimen lots, we believe the total number of arthropod specimens in North America would exceed 1 billion specimens (Derek Sikes, pers. Comm.). The overall pattern of records and diversity shows that compared to plants and especially

vertebrates, arthropod records are much lower for North America compared to their diversity (Table 1).

Table 1 presents summary statistics for digitization and species diversity for North American arthropod, plant, and chordate collections. The absolute number of digitized data records presented in GBIF is comparable for each group. However, the proportion of all North American arthropod specimens that have a digitized record is less than 5%, whereas that proportion is 15% for plants and higher for chordates. Moreover, because the total number of estimated arthropod species in North America is much greater than chordates and plants combined, the average number of specimens digitized per arthropod species (n=97) lags significantly behind both plants (n=404) and especially chordates (n=2,584).

In addition, GBIF currently serves some 330 million non-specimen-based records (e.g., eBIRD, (Sullivan et al. 2009)) and image-only records (e.g., iNaturalist, (Nugent 2018)) for chordates, which is nearly two orders of magnitude more than for plants and arthropods. In this regard, we also note that the Botanical Information and Ecology Network (BIEN) holds over 100 million observational records for New World plants (Enquist et al. 2016). In contrast, North American arthropods are only recently gaining traction in this arena, primarily due to citizen science initiatives such as iNaturalist, BugGuide.net, and other efforts focused on Lepidoptera (e.g., Butterflynet, Pollardbase) and Odonata (e.g., Xerces Society Dragonfly Pond Watch Project).

The Grand Digitization Challenge for North American Arthropod Collections

Given that North American collections hold approximately 300 million specimens, on what timeframe can we expect there to be a digital record available for each of those specimens? Figure 2 provides a visual representation of this "grand challenge." Our analyses indicate that some 2 million new digitized records are being produced annually from specimen labels, but as promising as this ongoing rate may be for generating large amounts of biodiversity data, there are still more than 280 million specimens remaining to be digitized. As a whole, we are currently not even transcribing enough specimen labels to keep up with new specimen acquisitions. A four-fold increase in our transcription rates is needed to capture label data for most specimens by mid-century (2050), assuming a 1% annual growth rate in specimen holdings.

The majority of the 223 collections and 300 million specimens in North America are located in the United States, although Canada and Mexico have representative holdings for their respective countries (Figure 3, Table S1). Canada has at least 17 collections and 32 million specimens, with the Canadian National Collection in Ottawa, Ontario curating 17 million of those specimens. The National Autonomous University of Mexico (UNAM) houses three million arthropod specimens, and its holdings comprise 97% of all estimated Mexican specimens in the country (but only seven other major collections were identified in Mexico). There are no published estimates for the number of arthropod species occuring in Mexico. However, some data are available for select groups such as the Arctiini (Lepidoptera: Noctuoidea: Erebidae: Arctiinae). In the United States and Canada, there are 237 species described in this tribe (Lafontaine and Schmidt, 2010) but over 385 species occur in Mexico (Diaz, 1996), which represents a 62% greater species diversity in Mexico. If co-occurring species are removed, about twice as many Arctiini occur in Mexico (n=289) compared to United States and Canada (n=141). These estimates are similar to a recent study demonstrating that vascular plant diversity is approximately 49% greater in Mexico compared to Canada and the United States

(Ulloa et al., 2017; despite the fact that Mexico contains only about 10% of the land area of Canada and the United States combined).
Given its greater projected arthropod diversity, Mexico would need to increase its specimen holdings 60-fold to generate a corpus of specimens comparable to that of collections in the United States and Canada. In terms of digitization progress, Mexico has conducted a major effort via CONABIO that resulted in 33% of their existing specimen labels being transcribed. This is a much greater proportion than either Canada (3%) or the United States (6%) has achieved to date.

The ADBC Initiative

Historically, individual taxonomists or ecologists working on a specific arthropod species and/or region conducted most digitization efforts, and those data were rarely shared. In just the past decade, the entomological community has made great strides in digitizing specimens and sharing those results (Figure 4). This effort has benefitted enormously from The National Science Foundation's Advancing Digitization of Biodiversity Collections (ADBC) program (iDigBio, 2019). ADBC began in 2011 and runs through 2021. More broadly, ADBC is enhancing and expanding the national resource of digital data that documents biological and paleontological collections, and is advancing scientific knowledge by improving access to digitized information (Nelson & Ellis 2018; Page et al. 2015).
The ADBC program has also promoted the development of a strong national investment in curation of the physical objects in scientific collections, and it contributes vitally to scientific research and technology interests in the United States. For arthropods, the impact of the ADBC program has been transformational from its inception, with the number of publicly available records having grown exponentially. Direct ADBC funding for digitization has produced about six million digitized records, and ADBC has indirectly spurred other collections to digitize their holdings. The NSF Collections in Support of Biological Research (CSBR) program has also emphasized digitization in its more recently funded CSBR awards.
The ADBC program has funded four Thematic Collections Networks (TCN) based on extant arthropods: InvertNet, Tri-Trophic, SCAN, and LepNet, with an additional TCN focused on invertebrates (InvertEbase) and an invasive species TCN that includes arthropods. The current TCN emphasis is on capturing descriptive data from specimen labels. However, collections are beginning to generate other data, such as geography, environmental habitat, phenology, associated organisms, collector field notes, and tissues and molecular data from specimens, which represent a rich biodiversity resource.
To expand on the recent ADBC efforts, we categorized North American collections into three groups based on digitization effort: (1) digitization not yet initiated; (2) records contributed to iDigBio, but no active digitization program in place; or (3) records contributed to iDiBio and with an active digitization program (Figure 5). We distinguished the latter two categories by whether there was an existing GBIF IPT (Integrated Publishing Toolkit) as an endpoint serving Darwin Core Archive data. It is encouraging that collections with active digitization programs account for 68% of the specimens in US collections, and that smaller collections that have not yet contributed data to public portals only account for 7% of collections. However, this underscores the need to extend digitization practices to smaller collections, because smaller collections are focal points for mentoring students who contribute to the national workforce. A major challenge

will be sustaining activities begun by ADBC activities once funding for the program ceases in 2021, such that collections can continue to integrate digitization into their everyday workflows.

Collection Holdings: Are We Meeting Research Data Needs?

It has been 28 years since Miller (1991) conducted the first and only comprehensive review of the 26 largest entomological collections at the time in the United States and Canada. The Miller review emerged from a 1988 meeting of the Association of Systematics Collections (ASC) that sought to address the capacity of systematics collections to increase research productivity, and proposed where national resources should be invested. As a measure of sustainability, the 26 collections in the Miller study have shown a steady 1% annual growth in the number of specimens, and the relative ranks of the collections have likewise remained rather stable (Figure 6, Table S2). We lack comparable statistics for the other 197 collections we surveyed in North America, but there are now 25 collections that house more specimens in 2018 than the 26th largest collection did in 1991 **(see** Table S1). Entomology collections in North America generally appear to be growing in the last ~30 years.

Are we collecting enough specimens?

North American collections have continued to grow three decades since Miller (1991) published his seminal paper, but we can still ask whether we are collecting enough. Securing sufficient resources to store and maintain specimens, and the steady 1% annual growth in specimen acquisition no doubt adds to the backlog of specimens needing to be digitized. Furthermore, it is becoming increasingly difficult to justify financial and personnel support for collections without making specimen data fully available to researchers and educators. With the exception of a few dedicated funding programs at NSF and the Institute of Museum and Library Services (IMLS), digitization has been a largely unfunded mandate for most institutions, adding significant budgetary pressure (Blagoderov et al. 2012; Heidorn 2011; Poole 2010). Global change impacts have elevated the urgency to develop regional to continental strategies for reaching appropriate targets for specimen holdings (Sánchez-Bayo & Wyckhuys 2019).

Will a projected 1% annual increase in specimen holdings meet expected future data needs? Are there enough arthropod specimens available now in collections for biodiversity-related research? We know that there are unmet research needs for specimen data (Kharouba et al. 2018), but it is difficult to grasp how acceptable the existing 300 million arthropod specimens are for meeting needs unless we continue to digitize specimens.

It is useful to compare efforts to digitize North American arthropods with that for vertebrates (see (Guralnick & Constable 2010)). Table 1 indicates that the average number of specimens digitized per arthropod species is 97, compared to 2,584 for chordate species, a 26-fold difference. We suggest that arthropod collections aim high and seek to digitize 2,500 records per species, to match efforts for chordates. We are not suggesting that 2,500 records are required for every arthropod species to address every question. Depending on the nature of the question, only a fraction of all available records may be appropriate (Piel, 2018; Veiga et al., 2017; Sikes et al., 2016; Ferro and Flick, 2015), and future analyses should provide more refined per species digitization targets (Lobo et al., 2018; Pelletier et al., 2018) once more digitized arthropod records become available.

We predict that to have a comparable corpus of arthropod data relative to chordates for North America, collections would need more than 360 million specimens to address data needs (Figure 7). This assumes that 60% (181 million) of the current 300 million specimens in arthropod collections are from North America, which may be an overestimate (but freshwater mollusk collections are estimated to be 60% for Canada and the United States; (Sierwald et al. 2018; Solem 1975)). The current rate of new specimen acquisition is insufficient, and even a doubling of the existing rate means that the target of 360 million would not be achieved until 2050. That target would be reached in 2047 if the overall rate of specimen acquisition were increased by 2.5% per year, by 2042 if it were increased to 3% annually and by 2030 if it were increased by 6% per year (Figure 7).

Two reasons to aim for 2,500 digitized records per arthropod species are taxonomic skew and spatial bias in digitized records. The average number of digitized records per North American arthropod species is 97 (Table 1). However, less than 15% of all 142,800 species have that many records, and only 0.1% have over 2,500 records. The most recorded species is *Bombus bifarius* (Cresson), a common bumblebee in western North America, with over 26,000 records. Even still, at its northern (Alaska) and southern (Arizona, Nevada and New Mexico) limits of this species' range, large gaps are present where there are few or no data records in areas they likely occur. This underscores that data bias can occur for even heavily sampled species (Ruete 2015). Moreover, many distribution maps for arthropod species (and other taxa) are incomplete and biased due to an overrepresentation of localities favored by collectors (e.g., roads, popular landmarks), in regions of otherwise more broadly suitable habitat. In addition to spatial bias, historical degradation of locality records is a major challenge (e.g., geopolitical name changes or imprecisely described localities; (Bartomeus et al. 2018)). One useful effort would be to resample for species that either have reliable historic records, and/or have the most vulnerable habitats that are either experiencing change or are predicted to change.

Assessing what is an adequate number of specimens has been initiated for two arthropod Thematic Collections Networks (SCAN, LepNet). Taxa being targeted range from individual species of conservation concern (e.g., Poweshiek Skipperling, *Oarisma poweshiek* (Parker);(Belitz et al. 2018)) to all Puerto Rican Lepidoptera that are susceptible to hurricanes (LepNet, 2019). In the case of *O. poweshiek*, it was determined that there were adequate numbers of existing specimens and observational records. For the assessment of Puerto Rican Lepidoptera, this prompted the launch of a longer-term inventory to obtain more complete collections of all Lepidoptera (Catherine Hulsof, pers. comm.). It is possible to provide reasonable running estimates for most North American species that provides basic metrics such as number of occurrences through time documented in suitable habitat or range. These can be used to guide individual species studies to target likely areas where species occur but have not been documented or resample historic areas to confirm their presence. The data for groups of species can be integrated into a more strategic plan to direct future sampling campaigns.

US Collections by Holding Size

Published reviews of natural history collections have focused on the collections with the largest specimen holdings (Dunnum et al. 2017; Miller 1991; Short et al. 2018; Sierwald et al. 2018; Singer et al. 2018). Here, we consider all collection sizes for the three North American countries, with a focus on the United States because it has more data that are readily available. We summarize basic characteristics of Tier 1 (largest) through Tier 4 (smallest) collections in

the US, including the number of collections, number of specimens, the percentage of collections that have initiated digitization, and the percentage of specimens that have had their labels transcribed for collections that are digitizing (Figure 8**)**. As expected, most collections are smaller (Tiers 3-4) although the absolute number of specimens is concentrated in larger collections (Tier 1). Small collections may face challenges in initiating digitization, but once begun, they processed a far greater percentage of their holdings than large collections. This suggests that NSF ADBC funding has been effective in promoting digitization across collections, but has not had as large an impact on the largest collections, where most specimens are located.

Table 2 shows additional metrics as a function of collection size. A general concern with the NSF ADBC program was whether smaller collections could adequately image specimens, provide digitized specimen data with species-level identifications, and properly georeference localities. We found relatively few significant differences in statistics among Tiers, although smaller collections appeared more effective in imaging, and small to intermediate sized collections more effective in identifications and georeferencing. We expected larger collections to have more global taxonomic and geographic coverage. To assess this, we measured the percentages of (a) non-North American records, (b) number of countries or large regional areas or islands, (c) total number of species recorded, and (d) the average distance of specimens from the collection itself. We predicted that smaller collections would have a strong regional focus and so we quantified (e) the percentage of specimens taken within a 50 km radius of the collection as a metric for a regional focus, and (f) the average rank collecting for each collection within the 50 km radius. These metrics supported our expectations, underscoring a more global taxonomic and geographic focus with increasing collection size. Distance from collection indicated a decreasing regional focus from Tier 1 to Tier 4 collections, although all collections had significant regional representation. The closest collection was almost always ranked first for having specimens from within 50 km of the collection. The only discrepancies occurred when two or more collections were physically near each other (e.g., Essig Museum in Berkeley, CA and the California Academy of Sciences in San Francisco, CA), or in a few Tier 1 collections (e.g., San Diego University, CA) where holdings strongly reflected a curator's research interest in taxa distributed outside of North America.

Possibly the most important metric regarding digitization was the number of "historical" records, which we defined as specimens collected prior to 1965, because these specimens represent perhaps the only direct evidence for pre-global change impacts (more fine-grained analysis of temporal patterns are underway; Cobb et al., unpubl. data). Our results show that large collections had more "historical" records than smaller ones (Figure 9), and that there are at least 32 million "historical" specimens in US collections that can be used to assess global change impacts on arthropods. This is encouraging but presents a challenge because specimens are typically not separated by sampling year in collections, and hence cannot be readily targeted for digitization. The typical practice for digitization is to digitize all specimens in a drawer, as it is extremely inefficient to digitize a fraction of specimens in a drawer or unit tray. Following Allan et al. (2019), we believe it is important to target special collections of historic importance and develop more effective ways to increase the overall efficiency of digitization.

Discussion

Moving Forward: Challenges and Opportunities

Our review is the first to provide a modern comprehensive assessment of arthropod collections in North America, and examine trends in the acquisition of new specimens and digitization of existing specimens. Both are important to address national/global needs for biodiversity data, and to initiate and promote collaborative networks among North American collections (organizations such as the Entomological Collections Network, CONABIO, and Canadensys already serve in this capacity and are well positioned to collaborate). Below we summarize key points of our findings, and propose actions needed to mobilize more collections-based arthropod data, to maintain the transformational effort initiated by the NSF ADBC program.

Increasing Specimen Holdings

We suggest that North American collections combined should increase the current holdings of North American arthropod specimens by at least an additional 100 million specimens by 2045 to marshal sufficient data to address global change impacts at the species level. This projection is based on the fact that less than 5% of all arthropod specimens in collections and only 0.1% of all arthropod species in collections are represented by species that have 2,500 digitized records/species – the average number of records/species digitized to date for North American chordate species. One hundred million specimens is a rough estimate that will have to be refined, but GA assessments should be done at the species level for priority arthropod taxa as we increase digitized records from collections, and develop research coordination networks to help guide and prioritize future surveys and digitization.

If we use estimates required for species distribution models, the expected standard for adequacy is growing, especially for species that occur over environmental gradients (Araújo et al. 2019). Thus, the target number of 100 million arthropod specimens may be an underestimate, given that 40% of the records in US collections are for specimens outside North America.

Increasing Digitization Efforts

We estimate that data label transcription rates will need to increase by at least four-fold if the rate of new specimen acquisition increases to 3% per year. This goal may be achievable if robotic technologies (e.g., Beyond the B) can be implemented at just Tier 4 collections. During the NSF ADBC funding years, a number of collections developed protocols for mass digitization of newly obtained material that are much more efficient than digitization of specimens already integrated into collections. Because Tier 1-2 collections only account for 6% of specimens in North American collections, they will not directly impact the total number of records, but they will have a significant effect on filling in regional gaps and/or focusing on specific arthropod taxa, and they are important for recruiting new biodiversity researchers.

Citizen Science and Computer-Aided Identification

To what degree can citizen science efforts help address the burgeoning arthropod data needs? Approximately 10% of arthropod species are thought to be identifiable to species using an image, date and geographic point location (http://www.lep-net.org/?page_id=25). As smartphone cameras improve, reference image databases expand, and citizen science programs like iNaturalist and FieldGuide continue to grow, we expect this to motivate biodiversity researchers to consider utilizing field images to augment physical specimens. Images are

currently accepted by GBIF as machine observations and along with human observations comprise the vast majority of GBIF records. The primary concern is that there is no physical specimen to confirm, and the vetting process is not as rigorous as desired. To date, records provided by iNaturalist to SCAN are primarily for those groups that are generally well known to entomologists. These include most species of Orthoptera, Odonata, and many Lepidoptera, along with specific taxa from other orders (e.g., Coccinellidae). Other arthropod orders (e.g., Araneae) still need to be evaluated to determine the degree to which species-level identifications can be obtained from images. Additionally, with the further genetic information on cryptic species (Miller et al. 2016) may identify more taxa that require more than images to obtain species-level identifications. Using images for identification will significantly help fill current gaps in arthropod data records, and occurrence records do not generally need to be transcribed from images (since modern phone cameras provide coordinate data). Heberling & Isaac (2018) list a suite of variables that can be captured by images of plants that are not typically available from herbarium specimens (e.g., color, biotic associations, habitat). The same is true for arthropods. All arthropods stored in alcohol or collected in ethyl acetate can experience color fading, and specimens left in sunlight or under fluorescent lighting can also lose their color. Host plant associations are typically not recorded, and if they are recorded, the plant specimen is usually not submitted as a corollary herbarium specimen. Computer-aided identification accuracy is increasing exponentially, with the primary limitation being the lack of training images for neural networks (Schuettpelz et al. 2017). Although data associated with specimens (images, genetics, observations) can help augment arthropod biodiversity data needs, they will never replace whole-specimen repositories.

Coordination among North American Countries

Although Mexico has made the greatest strides in digitization progress ~~(33% of their specimen labels are transcribed)~~, the 3 million specimens in Mexican collections remains low given that there are likely over 50,000 arthropod species in Mexico. Unlike the US and Canada, there are significant Mexican specimen holdings in institutions located in countries outside of Mexico. Many US taxa extend into Mexico, but the available data records often stop at the border (see Figure 10). There should be additional cross-country network development, (but note collaborative informal networks such as the Madrean Biodiversity Project that hosts various expeditions to northern Mexico; (Gottfried et al. 2013)).

Specimen holdings in Canadian collections are primarily of specimens from Canada and the northern US, and total around 32 million specimens. To date, Canada has recorded 20% of species diversity than the US but northern Canada, which harbors unique ecological habitats, are facing destruction and the remainder of the country may likely experience dramatic ecosystem conversion. The focus on the Arctic constitutes one of NSF's 10 Big Ideas for future research (https://www.nsf.gov/news/special_reports/big_ideas/). This NSF program should provide impetus for more specific planning and increased coordination among North American collections. Collections-based research will be important to these efforts, and there should be a North American effort to conduct repeated surveys (e.g., on a 3-5 year basis) to document the expected changes in the north.

Developing a Collections-based Network

Data collected during this review provide the basis for a permanent online repository similar to the Index Herbariorum for plant collections (Thiers 2015). We present a basic information framework in Table S1 necessary to establish such an online resource, and in which each collection could maintain its own data and integrate information from future work. We encourage the development of an "Index Entomologica" which could progressively add content such as sustainability scores for each collection based on criteria already established by the Index Herbariorum. The Entomological Collections Network (ECN; (Miller 1991)) acts as an umbrella organization for entomology collections to share best practices, and it could play a major role in supporting an Index Entomologica, along with other organizations such as the Society for the Preservation of Natural History Collections SPNHC (https://spnhc.biowikifarm.net/wiki). Although the ECN is primarily active in the United States, it also includes Canada and Mexico and is in a position to network further with entomology collections around the world. An Index Entomologica would be synergistic with the proposed “Extended Specimen Data” program that has emerged as the focus of future biodiversity efforts from the Biodiversity Collections Network (BCoN). Given that at least 90 million specimens in US collections are from countries outside of North America, the timing is ripe for North American collections to help build a global network with collaborations including e.g., iDigBio, GBIF, DISSCO, and SpeciesLink.

Next-Generation Collections

With a cohesive North American collection network in place, a new strategic plan should be implemented to augment the current rate of 1% annual growth in acquisition of new specimen numbers. Identifying gaps in taxonomic and geographic representation will lead to prioritization for collecting campaigns (e.g., the New Arctic). Existing collecting campaigns can also expand their efforts through temporary curation of by-catch samples to be shared with other researchers. The community as a whole should digitize and share by-catch samples (already implemented as part of the NEON ground dwelling carabid project). We have already seen a similar community effort in digitization campaigns in the LepNet TCN, where a group of over 50 collections focused their efforts on 3 target families of Lepidoptera, representing some of the most charismatic within the order (Papilionidae, Saturniidae, and Sphingidae).

NextGen collections is a new concept that has recently emerged from a national BCoN meeting (see themes outlined by Schindel & Cook (2018)). The primary focus is to promote integrated collections that include cross-phyla collections linked to environmental data gathered by deployable sensors. Collections are prioritized to address important social needs such as disease agents and pests. We fully support the NextGen concept, although the arthropod community still remains focused on filling taxonomic and regional gaps before this next step can be considered. Collecting data on associated taxa for key groups (herbivores, parasitoids, parasites, pollinators) and micro-environment data for other groups (detritivores, omnivores) are priorities. The resulting digitized data sets would promote more sophisticated and targeted efforts to better integrate data from collecting events.

NextGen collection practices will continue to arise in museums. For example, standard vocabularies will be developed for associated data denoting species associations (Poelen et al. 2014) and specimen traits, among others. It may not be feasible to employ robotic systems in all collections, but we can implement this technology through funding by programs that emerge from NSF’s 10 Big Ideas. Of the 10 Big Ideas, "Understanding the Rules of Life: Predicting the

Phenotype" is perhaps the most relevant because of the potential for coupling specimen-based research with targeted NextGen collections, and integration with ecological studies to understand how phenotypes evolve. Employing such techniques at just the 30 largest collections would allow the digitization of most specimens in North America in a shorter time than what we have estimated. Computer-aided identification tools can be deployed to help curators sort and identify specimens, and should be incorporated into NSF strategic planning as core programs emerge from NSF's 10 Big Ideas.

Conclusions

There are three major challenges and needs that remain for North American arthropod collections: (1) deploying effective strategies to integrate more specimens into collections; (2) improving of digitization workflows; and (3) better identification of societal needs for collection-based biodiversity information and conservation. To meet these challenges, there must be a strong call for a combination of technological development, financial and institutional resources needed to increase the capacity for needed specimens, and a better understanding of arthropods and their diversity. Increasing regional to global representation of arthropods will bring collections-based research to the forefront of addressing human impacts on our planet's biodiversity.

Acknowledgements

We thank Hojun Song and Jim Wooley (Texas A&M University) for providing their lists of collections, Larry Page (iDigBio, University of Florida) for providing lists of contacts. Katja Seltmann (University of California, Santa Barbara) helped conduct the initial set of analyses addressing digitization. Jesús Romero Napoles provided additional collections in Mexico. Scott Miller (Smithsonian Institute) kindly provided a review of the manuscript. We also thank the collaborating institutions that share data on the SCAN-LepNet portal, who provided data for this study. This work was supported in part by the following NSF grants: EF 1207371, DBI 1602081, DBI 1759966 to NSC; DBI 1600616 to LFG; DBI 1561448, DBI 1601957 to JMZ; DBI 1811897 to NJD; DBI 1601369 to AYK.

References

Allan EL, Livermore L, Price BW, Shchedrina O, and Smith VS. 2019. A novel automated mass digitisation workflow for natural history microscope slides. *Biodiversity Data Journal* 7.

Araújo MB, Anderson RP, Barbosa AM, Beale CM, Dormann CF, Early R, Garcia RA, Guisan A, Maiorano L, and Naimi B. 2019. Standards for distribution models in biodiversity assessments. *Science advances* 5:eaat4858.

Bakker FT. 2017. Herbarium genomics: skimming and plastomics from archival specimens. *Webbia* 72:35-45.

Bartomeus I, Stavert J, Ward D, and Aguado O. 2018. Historical collections as a tool for assessing the global pollination crisis. *Philosophical Transactions of the Royal Society B* 374:20170389.

Bebber DP, Carine MA, Wood JR, Wortley AH, Harris DJ, Prance GT, Davidse G, Paige J, Pennington TD, and Robson NK. 2010. Herbaria are a major frontier for species discovery. *Proceedings of the National Academy of Sciences* 107:22169-22171.

609 Belitz MW, Hendrick LK, Monfils MJ, Cuthrell DL, Marshall CJ, Kawahara AY, Cobb NS,
610 Zaspel JM, Horton AM, and Huber SL. 2018. Aggregated occurrence records of the
611 federally endangered Poweshiek skipperling (Oarisma poweshiek). *Biodiversity Data*
612 *Journal*.
613 Bell-Sakyi L, Darby A, Baylis M, and Makepeace BL. 2018. The Tick Cell Biobank: A global
614 resource for in vitro research on ticks, other arthropods and the pathogens they transmit.
615 *Ticks and tick-borne diseases* 9:1364-1371.
616 Besnard G, Christin P-A, Malé P-JG, Lhuillier E, Lauzeral C, Coissac E, and Vorontsova MS.
617 2014. From museums to genomics: old herbarium specimens shed light on a C3 to C4
618 transition. *Journal of experimental botany* 65:6711-6721.
619 Bieker VC, and Martin MD. 2018. Implications and future prospects for evolutionary analyses of
620 DNA in historical herbarium collections. *Botany Letters* 165:409-418.
621 Blagoderov V, Kitching IJ, Livermore L, Simonsen TJ, and Smith VS. 2012. No specimen left
622 behind: industrial scale digitization of natural history collections. *ZooKeys*:133.
623 Braun CE, and Wann GT. 2017. Historical Occurrence of White-Tailed Ptarmigan in Wyoming.
624 *Western North American Naturalist* 77:204-212.
625 Brooks DR, Hoberg EP, Boeger WA, Gardner SL, Galbreath KE, Herczeg D, Mejía-Madrid HH,
626 Rácz SE, and Dursahinhan AT. 2014. Finding them before they find us: informatics,
627 parasites, and environments in accelerating climate change. *Comparative Parasitology*
628 81:155-165.
629 Cho S, Epstein SW, Mitter K, Hamilton CA, Plotkin D, Mitter C, and Kawahara AY. 2016.
630 Preserving and vouchering butterflies and moths for large-scale museum-based molecular
631 research. *PeerJ* 4:e2160.
632 Cook JA, Edwards SV, Lacey EA, Guralnick RP, Soltis PS, Soltis DE, Welch CK, Bell KC,
633 Galbreath KE, and Himes C. 2014. Natural history collections as emerging resources for
634 innovative education. *Bioscience* 64:725-734.
635 Creley CM. 2016. Determining habitat suitability for the western gray squirrel and eastern gray
636 squirrel in California: Predicting future ranges with Maxent and ArcGIS. California State
637 University, Los Angeles.
638 Davis CC, Willis CG, Connolly B, Kelly C, and Ellison AM. 2015. Herbarium records are
639 reliable sources of phenological change driven by climate and provide novel insights into
640 species' phenological cueing mechanisms. *American Journal of Botany* 102:1599-1609.
641 Dunnum JL, Yanagihara R, Johnson KM, Armien B, Batsaikhan N, Morgan L, and Cook JA.
642 2017. Biospecimen repositories and integrated databases as critical infrastructure for
643 pathogen discovery and pathobiology research. *PLoS neglected tropical diseases*
644 11:e0005133.
645 Ellwood ER, Dunckel BA, Flemons P, Guralnick R, Nelson G, Newman G, Newman S, Paul D,
646 Riccardi G, and Rios N. 2015. Accelerating the digitization of biodiversity research
647 specimens through online public participation. *Bioscience* 65:383-396.
648 Ellwood ER, Kimberly P, Guralnick R, Flemons P, Love K, Ellis S, Allen JM, Best JH, Carter R,
649 and Chagnoux S. 2018. Worldwide Engagement for Digitizing Biocollections
650 (WeDigBio): The Biocollections Community's Citizen-Science Space on the Calendar.
651 *Bioscience* 68:112-124.
652 Ellwood ER, Monfils A, White L, Linton D, Douglas N, and Phillips M. 2019. Developing a
653 Data-Literate Workforce through BLUE: Biodiversity Literacy in Undergraduate
654 Education. *Biodiversity Information Science and Standards* 3:e37339.

Enquist BJ, Condit R, Peet RK, Schildhauer M, and Thiers BM. 2016. Cyberinfrastructure for an integrated botanical information network to investigate the ecological impacts of global climate change on plant biodiversity. PeerJ Preprints.

Evenhuis N, and Samuelson A. 2007. The insect and spider collections of the world.

Gottfried GJ, Ffolliott PF, Gebow BS, Eskew LG, and Collins LC. 2013. Merging science and management in a rapidly changing world: Biodiversity and management of the Madrean Archipelago III and 7th Conference on Research and Resource Management in the Southwestern Deserts; 2012 May 1-5; Tucson, AZ. *Proceedings RMRS-P-67 Fort Collins, CO: US Department of Agriculture, Forest Service, Rocky Mountain Research Station 593 p* 67.

Greve M, Lykke AM, Fagg CW, Gereau RE, Lewis GP, Marchant R, Marshall AR, Ndayishimiye J, Bogaert J, and Svenning J-C. 2016. Realising the potential of herbarium records for conservation biology. *South African Journal of Botany* 105:317-323.

Guralnick R, and Constable H. 2010. VertNet: creating a data-sharing community. *Bioscience* 60:258-259.

Hart R, Salick J, Ranjitkar S, and Xu J. 2014. Herbarium specimens show contrasting phenological responses to Himalayan climate. *Proceedings of the National Academy of Sciences* 111:10615-10619.

Heberling JM, and Isaac BL. 2018. iN aturalist as a tool to expand the research value of museum specimens. *Applications in plant sciences* 6:e01193.

Heidorn PB. 2011. Biodiversity informatics. *Bulletin of the American Society for Information Science and Technology* 37:38-44.

Janzen DH, and Hallwachs W. 2019. Perspective: Where might be many tropical insects? *Biological Conservation* 233:102-108.

Kharouba HM, Lewthwaite JM, Guralnick R, Kerr JT, and Vellend M. 2018. Using insect natural history collections to study global change impacts: challenges and opportunities. *Philosophical Transactions of the Royal Society B* 374:20170405.

Krishtalka L, and Humphrey PS. 2000. Can natural history museums capture the future? *Bioscience* 50:611-617.

Lacey EA, Hammond TT, Walsh RE, Bell KC, Edwards SV, Ellwood ER, Guralnick R, Ickert-Bond SM, Mast AR, and McCormack JE. 2017. Climate change, collections and the classroom: using big data to tackle big problems. *Evolution: Education and Outreach* 10:2.

Lawson S, Shell W, Lombard S, and Rehan S. 2018. Climatic variation across a latitudinal gradient affect phenology and group size, but not social complexity in small carpenter bees. *Insectes sociaux* 65:483-492.

Lister BC, and Garcia A. 2018. Climate-driven declines in arthropod abundance restructure a rainforest food web. *Proceedings of the National Academy of Sciences* 115:E10397-E10406.

McIntyre NE. 2000. Ecology of urban arthropods: a review and a call to action. *Annals of the Entomological Society of America* 93:825-835.

McLean N, Lawson CR, Leech DI, and van de Pol M. 2016. Predicting when climate-driven phenotypic change affects population dynamics. *Ecology Letters* 19:595-608.

Meineke EK, Davies TJ, Daru BH, and Davis CC. 2018. Biological collections for understanding biodiversity in the Anthropocene. The Royal Society.

Miller SE. 1991. Entomological collections in the United States and Canada. *American Entomologist* 37:77-84.
Miller SE, Hausmann A, Hallwachs W, and Janzen DH. 2016. Advancing taxonomy and bioinventories with DNA barcodes. *Philosophical Transactions of the Royal Society B: Biological Sciences* 371:20150339.
Monfils AK, Powers KE, Marshall CJ, Martine CT, Smith JF, and Prather LA. 2017. Natural history collections: teaching about biodiversity across time, space, and digital platforms. *Southeastern Naturalist* 16:47-58.
Nelson G, and Ellis S. 2018. The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B* 374:20170391.
Nufio CR, McGuire CR, Bowers MD, and Guralnick RP. 2010. Grasshopper community response to climatic change: variation along an elevational gradient. *PLoS One* 5:e12977.
Nugent J. 2018. iNaturalist: Citizen Science for 21st-Century Naturalists. *Science Scope* 41:12.
Page LM, MacFadden BJ, Fortes JA, Soltis PS, and Riccardi G. 2015. Digitization of biodiversity collections reveals biggest data on biodiversity. *Bioscience* 65:841-842.
Poelen JH, Simons JD, and Mungall CJ. 2014. Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecological Informatics* 24:148-159.
Poole PN. 2010. The Cost of Digitising Europe's Cultural Heritage A Report for the Comité des Sages of the European Commission.
Poss SG, and Collette BB. 1995. Second survey of fish collections in the United States and Canada. *Copeia*:48-70.
Primack RB, and Gallinat AS. 2017. Insights into grass phenology from herbarium specimens. *New Phytologist* 213:1567-1568.
Ratnasingham S, and Hebert PD. 2007. BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). *Molecular ecology notes* 7:355-364.
Ruete A. 2015. Displaying bias in sampling effort of data accessed from biodiversity databases using ignorance maps. *Biodiversity Data Journal*
Sánchez-Bayo F, and Wyckhuys KA. 2019. Worldwide decline of the entomofauna: A review of its drivers. *Biological Conservation* 232:8-27.
SCAN. 2019. The Symbiota Collections of Arthropods Network (SCAN) serves specimen occurrence records and images from North American arthropod collections.
Schindel DE, and Cook JA. 2018. The next generation of natural history collections. *PLoS biology* 16:e2006125.
Schmitt CJ, Cook JA, Zamudio KR, and Edwards SV. 2018. Museum specimens of terrestrial vertebrates are sensitive indicators of environmental change in the Anthropocene. *Philosophical Transactions of the Royal Society B* 374:20170387.
Schuettpelz E, Frandsen PB, Dikow RB, Brown A, Orli S, Peters M, Metallo A, Funk VA, and Dorr LJ. 2017. Applications of deep convolutional neural networks to digitized natural history collections. *Biodiversity Data Journal*.
Seltmann KC, Cobb NS, Gall LF, Bartlett CR, Basham MA, Betancourt I, Bills C, Brandt B, Brown RL, and Bundy C. 2017. LepNet: The Lepidoptera of North America Network. *Zootaxa* 4247:73-77.
Short AEZ, Dikow T, and Moreau CS. 2018. Entomological collections in the age of big data. *Annual Review of Entomology* 63:513-530.

Sierwald P, Bieler R, Shea EK, and Rosenberg G. 2018. Mobilizing mollusks: Status update on mollusk collections in the USA and Canada. *American Malacological Bulletin* 36:177-215.

Sikes DS, Copas K, Hirsch T, Longino JT, and Schigel D. 2016. On natural history collections, digitized and not: a response to Ferro and Flick. *ZooKeys*:145.

Singer RA, Love KJ, and Page LM. 2018. A survey of digitized data from US fish collections in the iDigBio data aggregator. *PLoS One* 13:e0207636.

Solem A. 1975. The Recent mollusk collection resources of North America. *The Velinger* 18:222-236.

Song H. 2019.

Spear DM, P[illegible]y GB, and Kaiser K. 2017. Citizen science as a tool for augmenting museum collection data from urban areas. *Frontiers in Ecology and Evolution* 5:86.

Ströbel B, Schmelzle S, Blüthgen N, and Heethoff M. 2018. An automated device for the digitization and 3D modelling of insects, combining extended-depth-of-field and all-side multi-view imaging. *ZooKeys*:1.

Sullivan BL, Wood CL, Iliff MJ, Bonney RE, Fink D, and Kelling S. 2009. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation* 142:2282-2292.

Thiers B. 2015. continuously updated]: Index Herbariorum: A global directory of public herbaria and associated staff. New York Botanical Garden's Virtual Herbarium. *Published at http://sweetgum nybg[illegible]g/science/ih/[last accessed 13 Jul 2016]*.

Thiers B, Mabee P, and Mon[illegible]s A. 2019. Extending US Biodiversity Collections to Address National Challenges. *Biodiversity Information Science and Standards* 3:e37225.

van de Kamp T, Cecilia A, dos Santos Rolo T, Vagovič P, Baumbach T, and Riedel A. 2015. Comparative thorax morphology of death-feigning flightless cryptorhynchine weevils (Coleoptera: Curculionidae) based on 3D reconstructions. *Arthropod structure & development* 44:509-523.

Watanabe ME. 2019. The Evolution of Natural History Collections: New research tools move specimens, data to center stage. *Bioscience* 69:163-169.

Weirauch C, Seltmann KC, Schuh RT, Schwartz MD, Johnson C, Feist MA, and Soltis PS. 2017. Areas of endemism in the Nearctic: a case study of 1339 species of Miridae (Insecta: Hemiptera) and their plant hosts. *Cladistics* 33:279-294.

Willis CG, Ellwood ER, Primack RB, Davis CC, Pearson KD, Gallinat AS, Yost JM, Nelson G, Mazer SJ, and Rossington NL. 2017. Old plants, new tricks: Phenological research using herbarium specimens. *Trends in Ecology & Evolution* 32:531-546.

Table 1(on next page)

Metrics for North American collections for Arthropoda, Chordata, and Plantae.

Species richness for Chordata estimated from (Dunnam et al 2018), for Plantae from (Ulloa et al., 2017) and for Arthropoda from Stork (2018). Data obtained from GBIF in January 2019.

1

	Arthropoda	Chordata	Plantae
# Species	142,800	4,424	34,109
# Specimen Records	13,788,159	11,430,528	13,787,883
# Non-Specimen Records	3,335,975	329,994,473	6,729,368
# Records/Species (Specimen Records)	97	2,584	404
# Records/Species (Non-Specimen records)	23	74,597	197

2

Table 2(on next page)

Summaries of metrics for digitized records from the four size Tier categories.

Standard error of means are provided where applicable.

1

	Tier Collection Size Categories				
Collection size categories	**Tier 1 < 0.1 million**	**Tier 2 0.1-1 million**	**Tier 3 1-3 million**	**Tier 4 >3 million**	**Trend**
Data Quality					
Georeferenced	60% (+11)	72% (+9)	72% (+8)	60% (+8)	**none**
Identified to species	51% (+8)	62% (+6)	70% (+6)	57% (+7)	**nonlinear**
Records with images	22% (+10)	19% (+8)	6% (+4)	11% (+6)	**down**
Regional to Global Metrics					
Non-North America records	15% (+7)	10% (+3)	20% (+6)	48% (+9)	**up**
# of Countries/major regions	69	61	197	355	**up**
Species per collection	631 (+258)	2,713 (+437)	4,451 (+1,353)	16,990 (+6,884)	**up**
Distance from Collection (km)	881 (+343)	621 (+146)	1,106 (+174)	2,850 (+725)	**up**
% of records (50 km radius)	85 (± 5)	63 (+5)	62 (+5)	43 (+7)	**down**
Mean rank (50km radius)	1 (± 0.0)	1 (± 0.0)	1.1 (+0.1)	1.5 (+0.2)	**None**

2
3

Figure 1

Map of North America showing the location of the arthropod collections included in the present study.

Alaska and Hawaii are shown as inserts in lower left (Guam not shown).

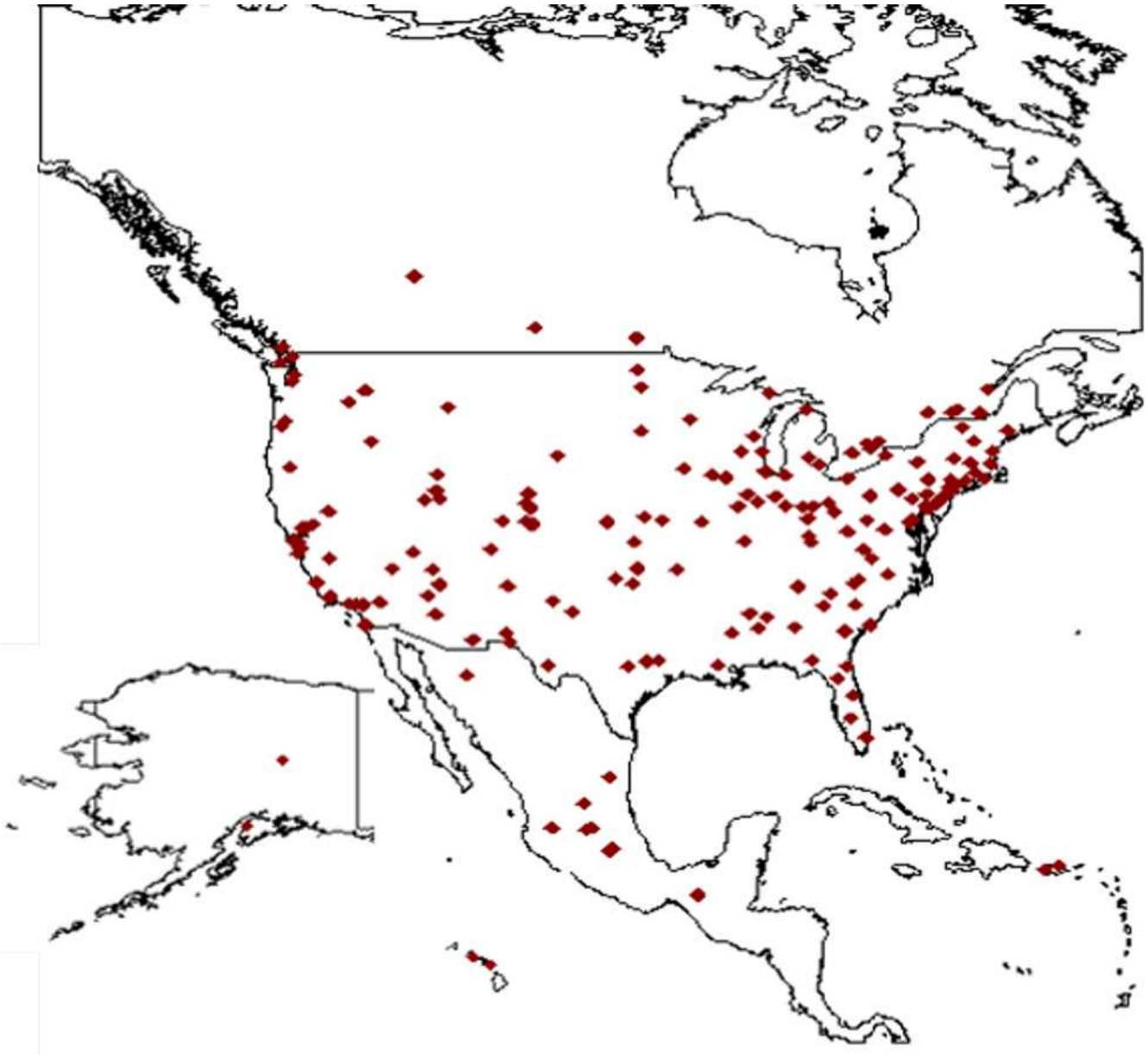

Figure 2

The grand challenge for North American arthropod collections.

A. Number of records of specimens digitized through 2018 (blue bar, in millions) and the total number of specimens in collections (green bar). B. Projections of ongoing acquisition rates for specimens, compared to rates of digitization.

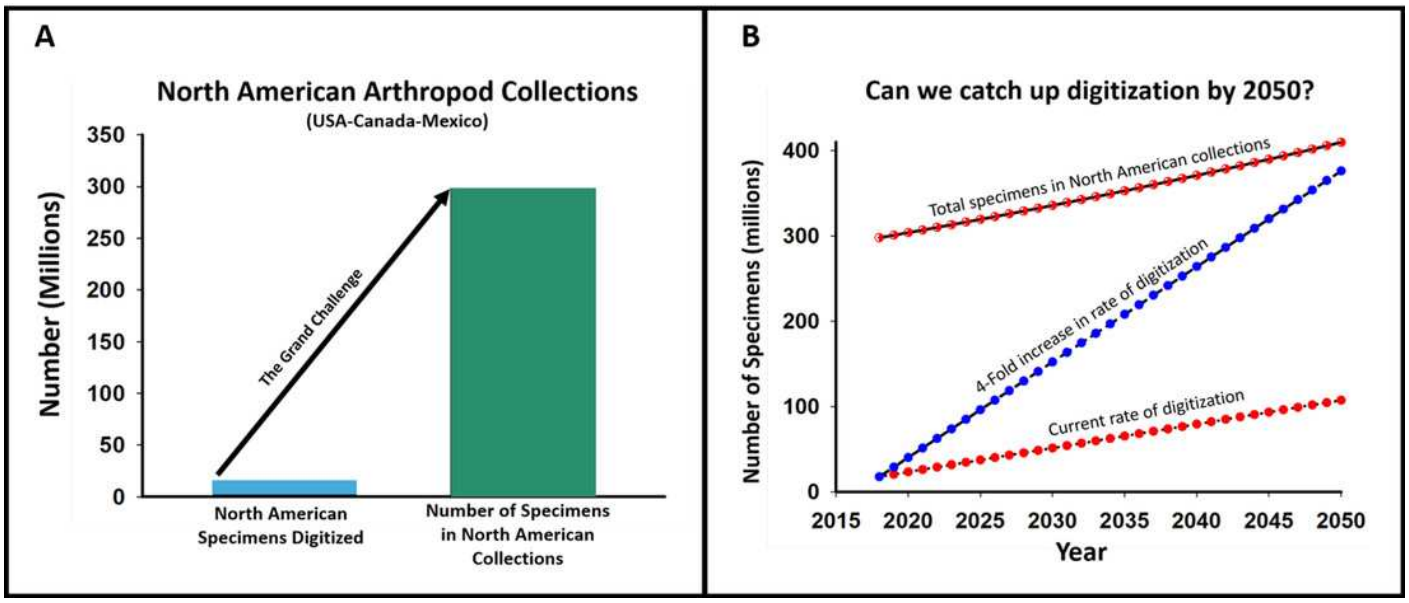

Figure 3

Number of arthropod collections (blue) and number of specimens (green) for North American collections.

The current percent of specimens whose label data have been transcribed is above each bar.

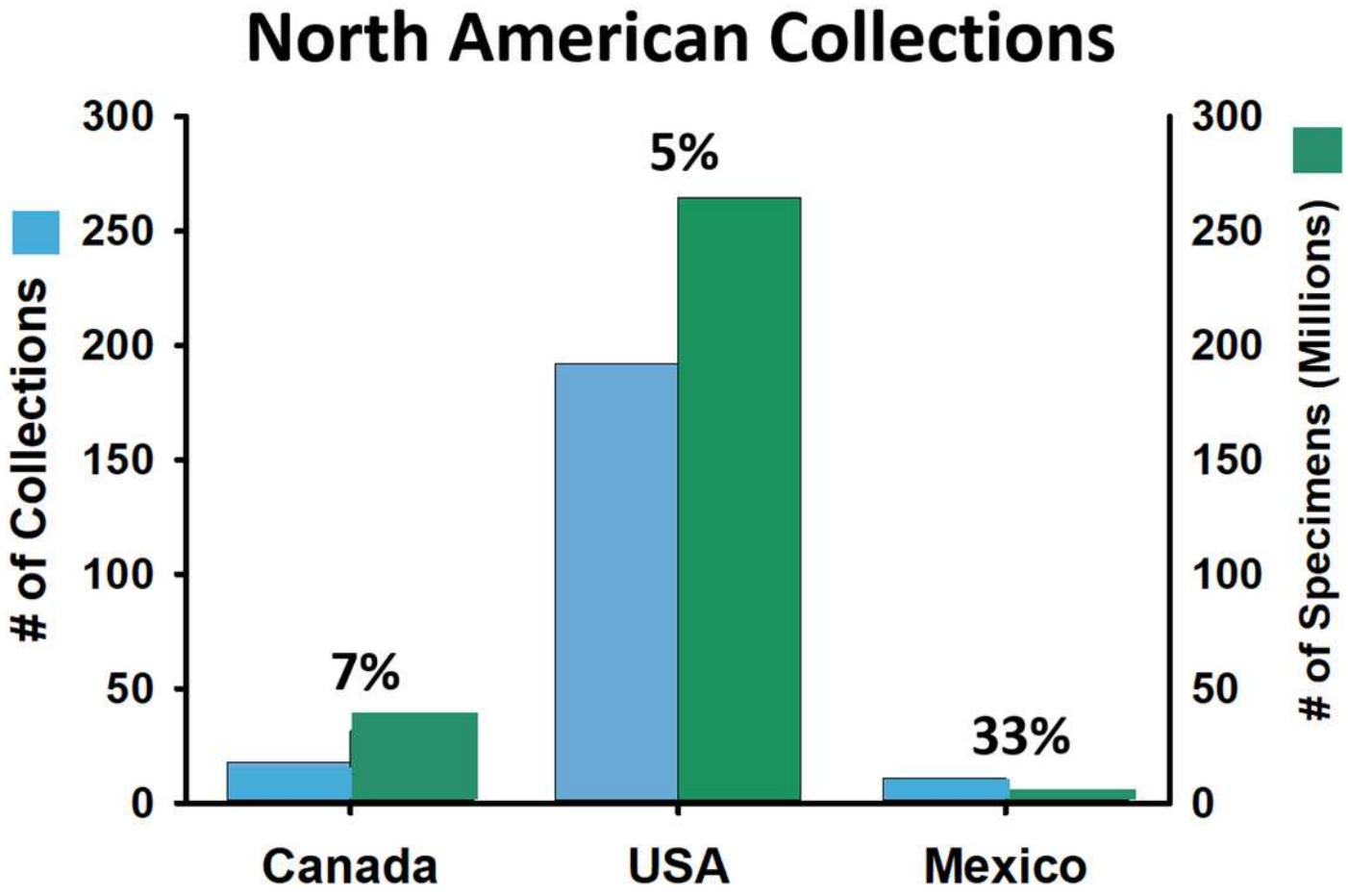

Figure 4

Number of digitized occurrence records for arthropod specimens from North American collections.

Estimates before 2010 are from Miller (1991), estimates since are from periodic queries of GBIF and SCAN.

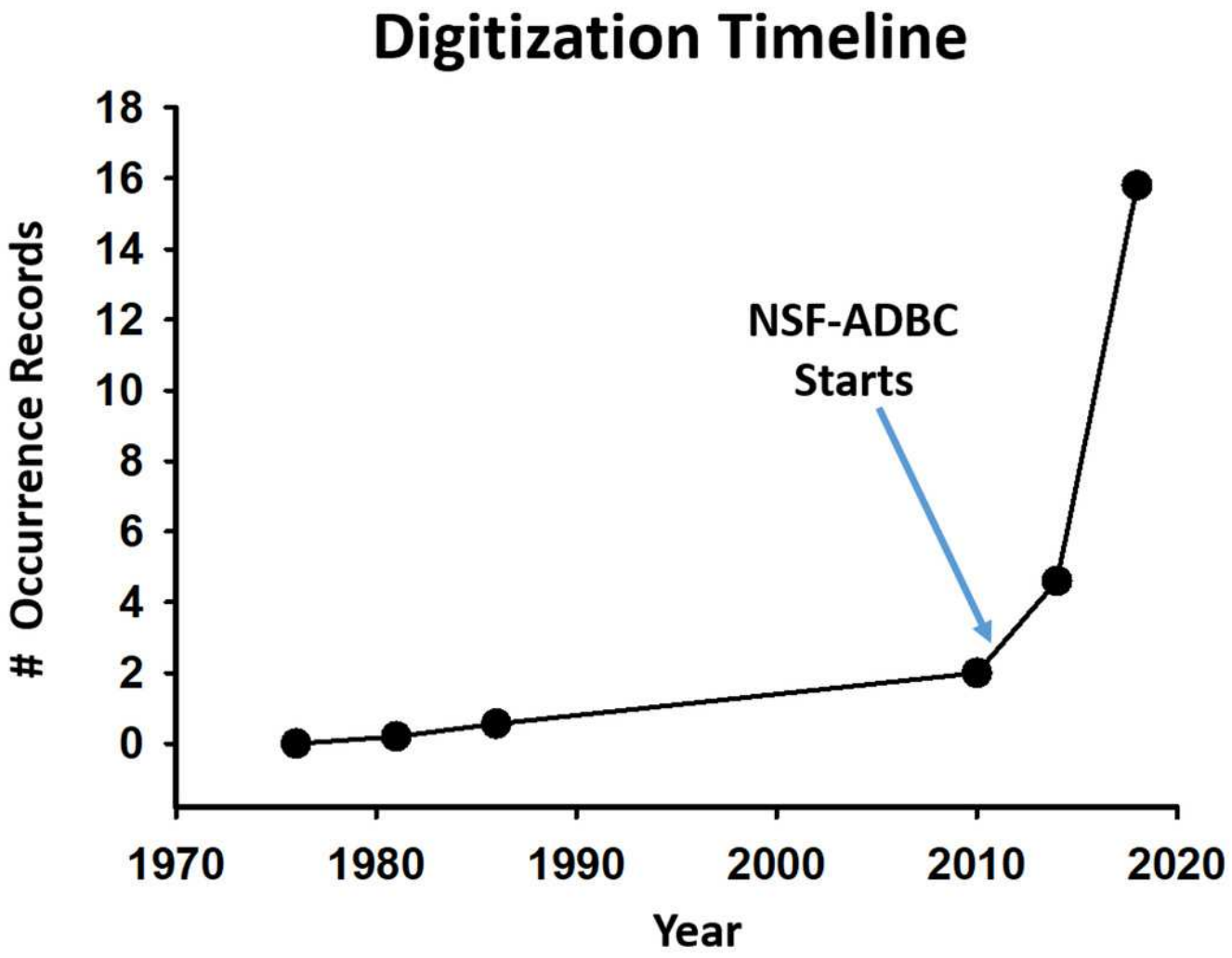

Figure 5

Number of US collections and percentage of US specimens.

Collections are arranged by degree of digitation effort; see text for elaboration of effort categories.

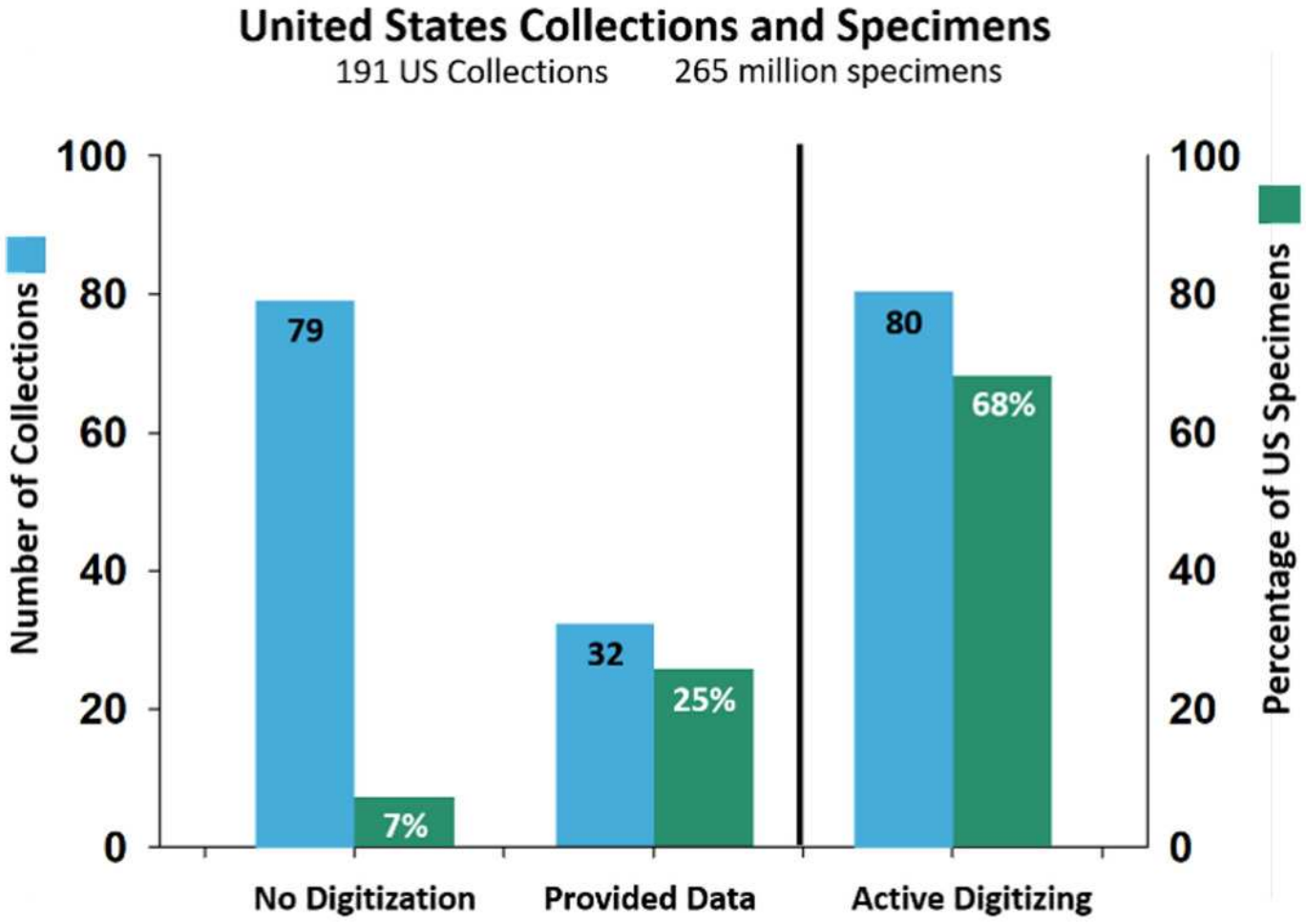

Figure 6

Growth in number of specimens at the 26 largest collections in Canada and the United States over three decades.

Estimates from 1980's tabulated by Miller (1991), 2018 estimates extracted from this review (Table S1).

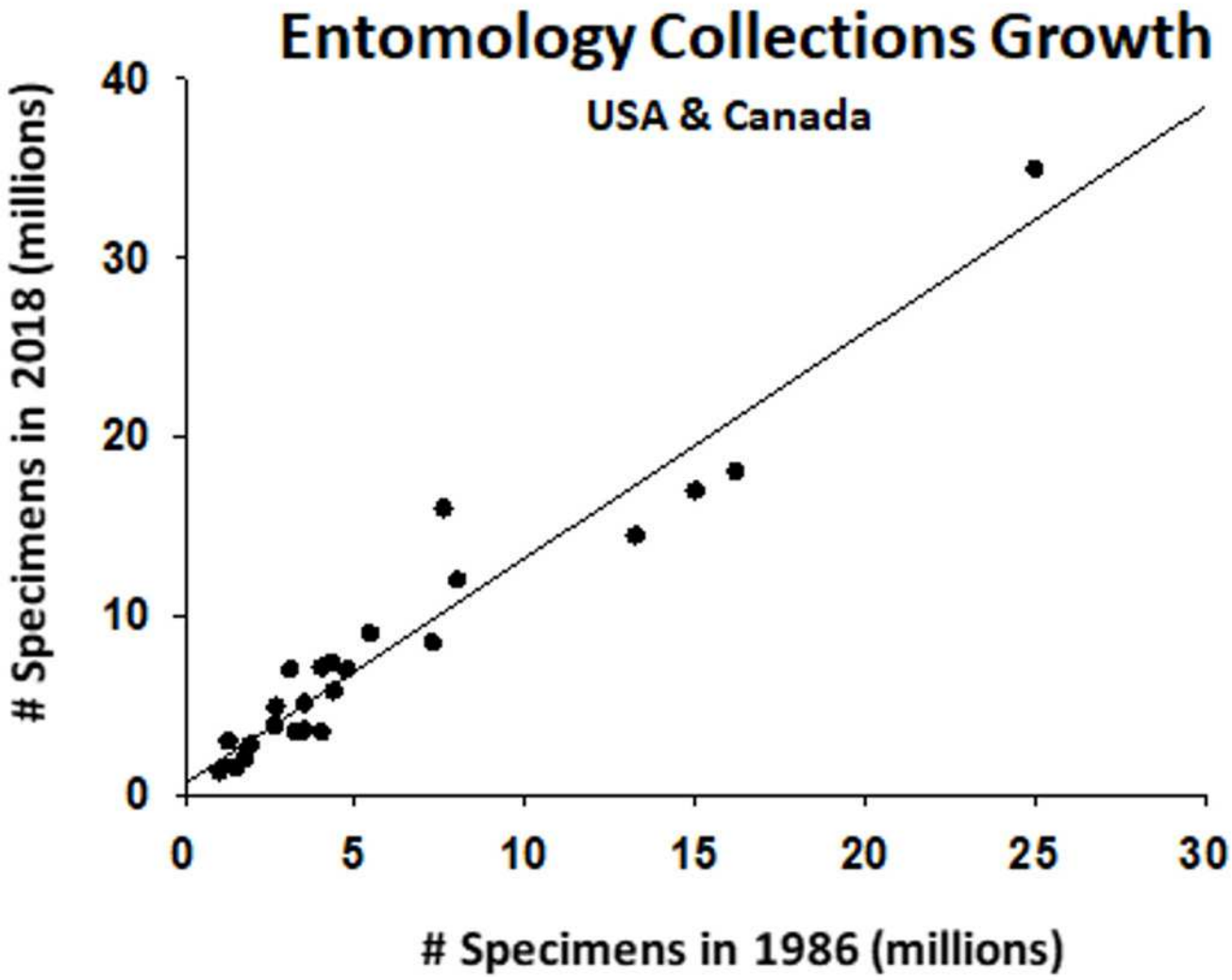

Figure 7

Projected growth in specimen numbers that would be required to meet data demands for biodiversity research.

Values expressed as percent increase in North American holdings for all collections in North America. Red circles indicate goals under the two trajectories.

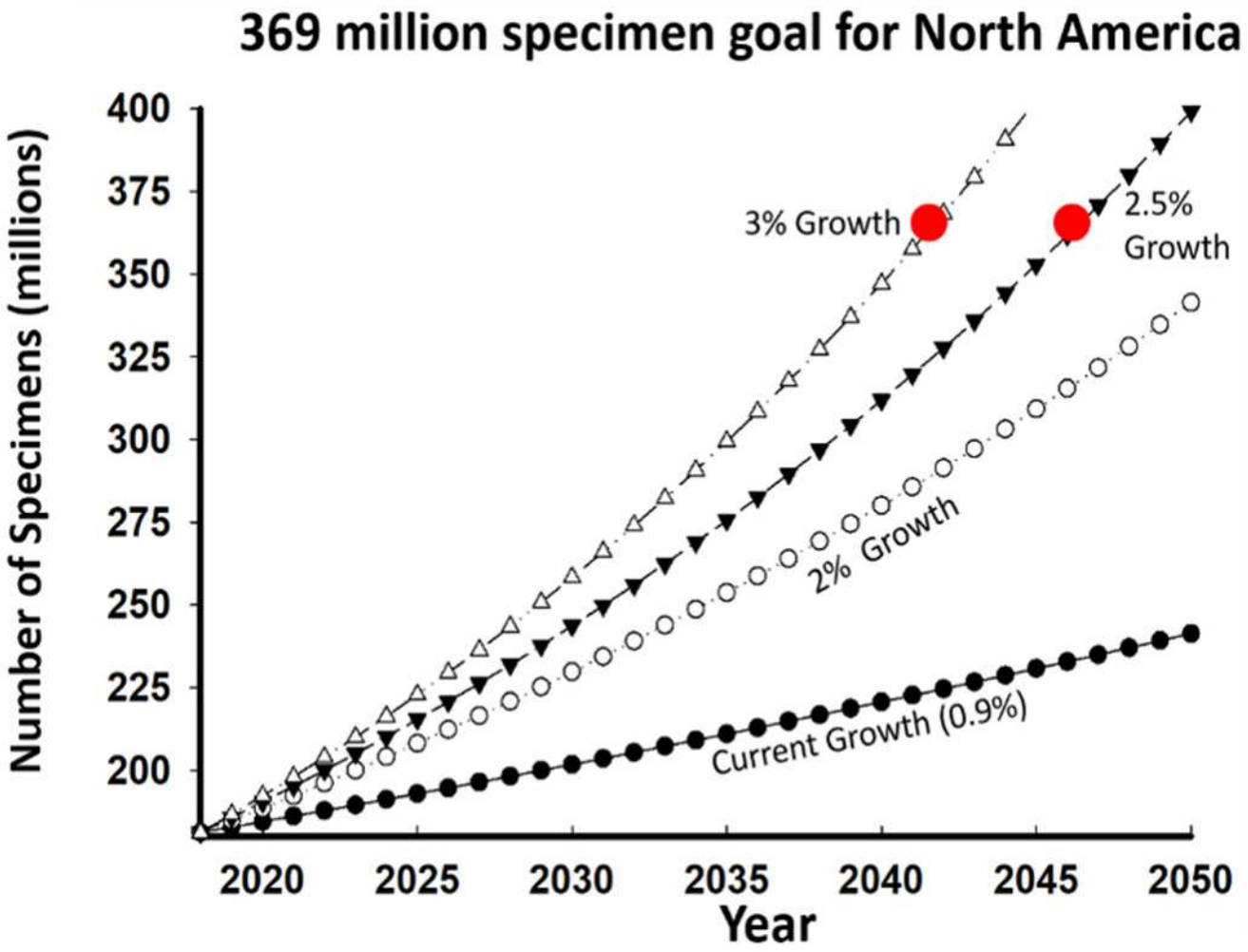

Figure 8

Attributes of 189 US collections arranged by size.

Tier 1: < 100,000 specimens, Tier 2: 100,000 to 1,000,000 specimens, Tier 3: 1,000,000 to 3,000,000 specimens, Tier 4: Over 3,000,000 specimens. Numbers within black bars either represent the numbers of collections (A, C) or percentage values for each Tier (B, D).

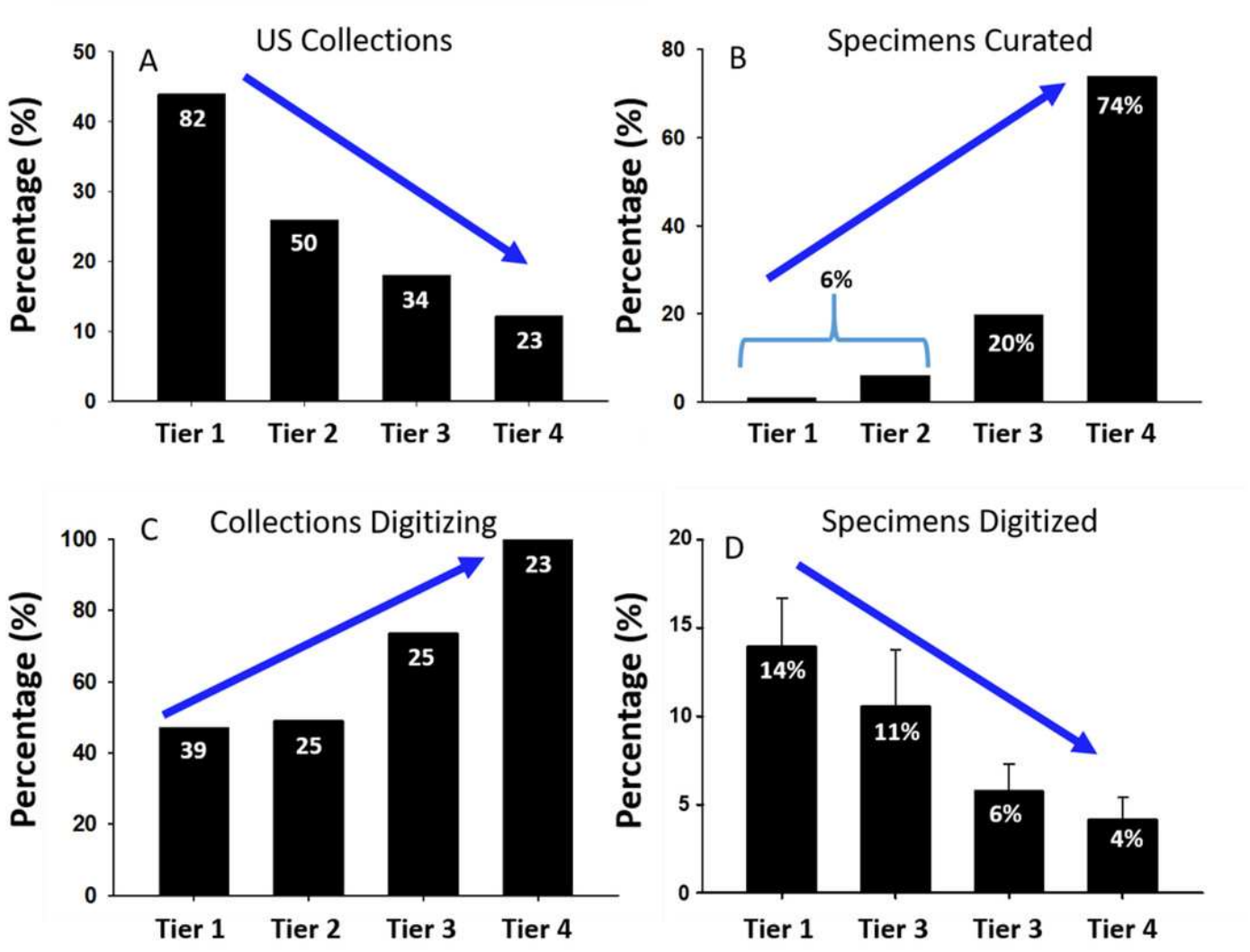

Figure 9

Estimates for numbers of specimens collected prior to 1965 in US collections.

Tier 4 collections hold the vast majority of “historical” specimens.

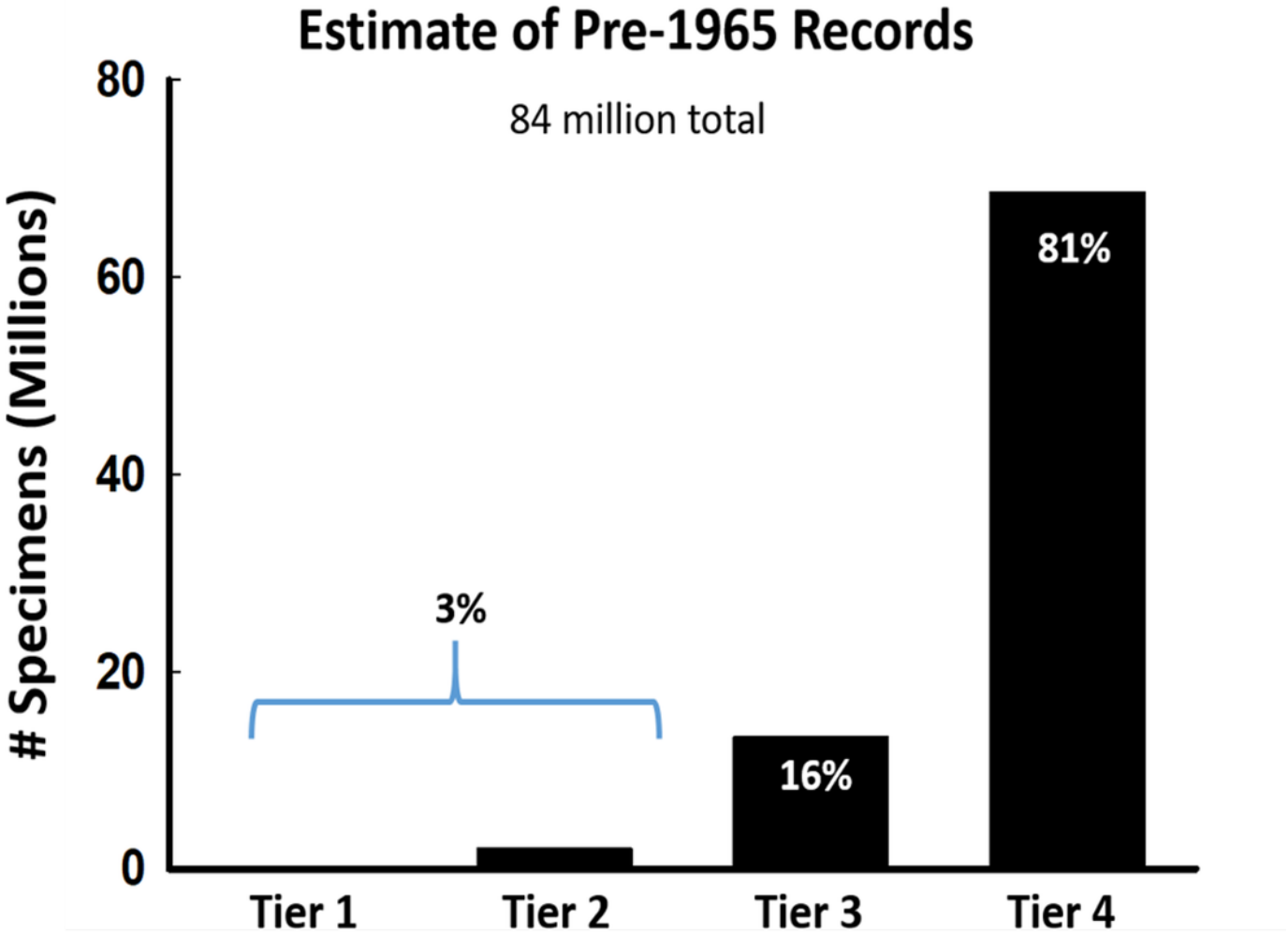

Figure 10

Heat maps showing distributions for *Lasius* (Formicidae) and *Bembidion* (Carabidae) from SCAN data

The dashed ellipses show a “border impact” where there is strong coverage in the US but almost no records in Mexico. Record density ranges from red (high) to green (low). Data derived from SCAN Spatial Module (heat map radius=1, blur=4).