The Changing Uses of Herbarium Data in an Era of Global Change: An Overview Using Automated Content Analysis

J. MASON HEBERLING, L. ALAN PRATHER, AND STEPHEN J. TONSOR

Widespread specimen digitization has greatly enhanced the use of herbarium data in scientific research. Publications using herbarium data have increased exponentially over the last century. Here, we review changing uses of herbaria through time with a computational text analysis of 13,702 articles from 1923 to 2017 that quantitatively complements traditional review approaches. Although maintaining its core contribution to taxonomic knowledge, herbarium use has diversified from a few dominant research topics a century ago (e.g., taxonomic notes, botanical history, local observations), with many topics only recently emerging (e.g., biodiversity informatics, global change biology, DNA analyses). Specimens are now appreciated as temporally and spatially extensive sources of genotypic, phenotypic, and biogeographic data. Specimens are increasingly used in ways that influence our ability to steward future biodiversity. As we enter the Anthropocene, herbaria have likewise entered a new era with enhanced scientific, educational, and societal relevance.

Keywords: automated content analysis, biological collections, herbarium, museum, specimen

Collected and curated by thousands of botanists over nearly five centuries, herbaria comprise an enormous, internationally shared scientific enterprise that archives the world's botanical and fungal diversity through time. Nearly 390 million specimens of preserved plants, fungi, algae, and related taxa currently reside in over 3000 herbaria across the world (figure 1; Thiers 2018). The users of herbaria, like those of other natural history collections, were historically dominated by taxonomists who focused on describing the species of the world, understanding how they were distributed across the continents, and elucidating their relationships (i.e., cataloging, describing, and organizing; Funk 2018). Herbarium specimens continue to serve core roles in taxonomy, floristics, and species identification, as well as as scientific vouchers (Funk et al. 2005) and in education (Cook et al. 2014, Monfils et al. 2017), but the growing number of reviews on modern uses of herbarium data strongly suggest that we have entered a distinctly new era for specimen use (e.g., Nualart et al. 2017, Bieker and Martin 2018, Carine et al. 2018, Lang et al. 2019, Meineke et al. 2018b).

Although the longstanding uses of herbaria remain important, the availability of new tools for data access and analysis has the capacity to expand the use of herbarium data beyond conventional research (Nelson and Ellis 2018), perhaps redefining the core functions of herbaria altogether (Heberling and Isaac 2017). For example, taxonomy and systematics have always been at the center of herbarium research, but methodological advances have revolutionized these studies, especially through extraction of genetic and even expression information from century old herbarium specimens (Buerki and Baker 2016). Similarly, recent developments in artificial intelligence are being applied to specimen identification on a large scale (Carranza-Rojas et al. 2017) and automated/semi-automated approaches for extracting phenotypic data from specimens are becoming common (high-throughput phenotyping; Gehan and Kellogg 2017). Along with the creation of global digital data repositories, methodological advances in statistics, computer science, and geography have transformed the representation of species distributions from comparatively coarse occurrence maps to predictions of niche suitability and biodiversity across scales (Soltis 2017).

Examination of recent papers suggests that the traditional foci of research utilizing herbarium specimens, although still important, now exist alongside a number of novel, unanticipated research uses (Heberling and Isaac 2017). Several recent reviews have focused on cutting-edge applications

BioScience 69: 812–822.
doi:10.1093/biosci/biz094 Advance Access publication 4 September 2019

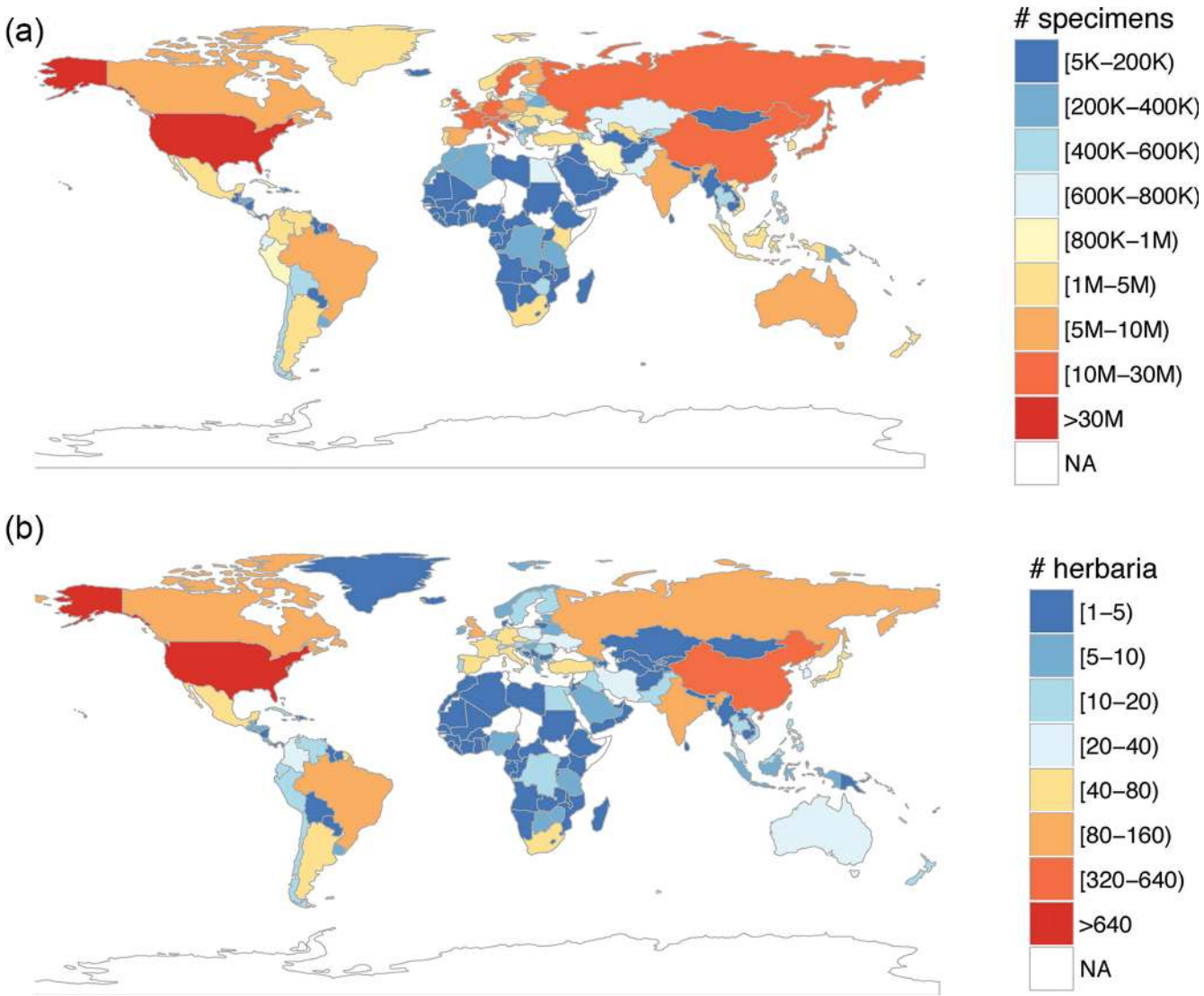

Figure 1. Country-level representation of (a) the number of specimens and (b) the number of active herbaria, as recorded in Index Herbariorum (data from Thiers 2018). Worldwide, there are an estimated 381,308,064 specimens residing in 2962 herbaria. Maps were made using the choroplethr package in R (Lamstein and Johnson 2017).

in single research areas, including evolution (Holmes et al. 2016), conservation biology (Nualart et al. 2017), global change biology (Lang et al. 2019, Meineke et al. 2018a), and environmental studies (Lavoie 2013). A systematic synthesis and categorization of the uses of herbarium specimen data and the relative frequencies of the categories through time is needed to place these trends in a broader context. A broad understanding of these trends will promote continued development for the next generation of herbarium use.

In this paper, we quantitatively review the scope, content, and trends in the herbarium-based literature over the last 125 years utilizing a recently developed machine-learning based method known as topic modeling (automated content analysis) for large-scale computational text analysis (Nunez-Mir et al. 2016). We gained insights from a combination of computational text analysis to identify and quantify trends and traditional literature review to interpret and contextualize these trends. Specifically, we asked: (1) Do herbarium specimens remain relevant in contemporary research? That is, are herbarium-based studies published at similar rates compared to other studies in the plant sciences? (2) What are the major scientific uses of herbarium specimen data? (3) How have the prevailing uses of herbarium data changed through time and how does this inform our prioritization of continued investment in herbarium-centered activities?

Automated content analysis of the herbarium-based literature

We took a broad approach to identify and compile published journal articles that explicitly mentioned herbarium

specimens or specimen-derived data in the title, abstract, and/or keywords. Literature searches were done using the topic field in Web of Science (Clarivate Analytics, formerly ISI) and the Elsevier Scopus database. The final dataset analyzed consisted of 13,702 herbarium-related articles, published from 1923 to 2017. See supplemental material for further details on literature searches.

Structural topic modeling

To gain a quantitative understanding of the major content and temporal trends represented in herbarium-based studies, we performed a form of large-scale computational text analysis called probabilistic topic modeling, also called automated content analysis. In brief, topic modeling is a powerful approach to synthesize an overwhelmingly large volume of literature, in which a set of texts are coded into meaningful topics through computer-assisted text processing and classification. These topics (sometimes called "concepts" or "themes") are defined based on word co-occurrence within and between texts and their prevalence across the entire body of text (corpus). The topics themselves can be user defined (supervised models) or can emerge inductively based on text patterns identified through computer algorithms (unsupervised). Topic models are particularly popular in the social sciences and humanities (Roberts et al. 2014), but only recently applied in ecology and evolutionary biology (Nunez-Mir et al. 2016).

In this review, we used an approach called structural topic modeling (STM; Roberts et al. 2014), which is derived from the latent Dirichlet allocation (LDA) topic model approach (Blei et al. 2003). STM is an unsupervised, "mixed-membership" model, wherein each word within a document is classified to a given topic and each document can include multiple topics. Each document is represented as a vector of topic proportions according to fractions of words assigned to a given topic. Topics are therefore defined by probability distributions for a vocabulary of words that group together. Because we were specifically interested in how the prevalence of herbarium-related topics change through time, we included publication year as a covariate. Because article titles often include focal words that are not repeated in the abstract or keywords, we combined article titles and abstracts for analysis. Analyses were performed using the stm package (Roberts et al. 2017) in R (R Core Team 2017).

Topic models (including the STM approach) have several strengths as tools for identification of topic sets in a body of literature (Nunez-Mir et al. 2016). First, as the number of documents and topics increases, manual inclusion of documents in a topic set becomes exceedingly difficult or impossible, whereas topic models can handle large numbers of publications and topics. Second, as an unsupervised approach, this method avoids potential bias based on the expectations of person(s) manually assessing text (i.e., using an *ex ante* definition of topics). Last, because topic definition is unsupervised, topic models allow for the emergence of unexpected research themes in a body of literature—that is, researchers can discover topics rather than assume them (Roberts et al. 2014). Therefore, topic models provide an efficient, transparent, replicable method well suited for systematic literature reviews. It is important to note that topic models only reveal underlying structure in the text by identifying groups of words that tend to co-occur. It is up to the researcher to interpret the meaning of these groups, that is, identify the topic meaning or subject matter.

Although the STM approach is computer automated, a solid understanding of the analyzed texts is required by the researcher to make model selection decisions and interpret the meaning of the resulting topics. The "optimal" number of topics modeled depends on prior research on the subject matter, the scope of goals or questions motivating the analysis, and the corpus itself (Farrell 2016). Modeling too few topics lumps otherwise meaningful topics into broad categories that may blur interpretation and modeling too many topics adds superfluous complexity and may result in many topics that lack substantive meaning. Following Farrell (2016), we approached model selection as a recursive process requiring expert qualitative evaluation of model results. We compared the output from a range of models that differed in the number of topics modeled. For each model, we closely read a subset of the abstracts that were most highly associated with each topic to assess cohesiveness and interpretative value and to assign a topic subject matter description to each topic. We then iteratively increased the number of topics until the increase in meaningful topics decreased relative to the number that had no clear meaning. We report results from the optimal 25-topic model, discarding 3 topics that lack interpretive value. See supplemental materials for additional details on topic modeling methods, including model selection and validation.

Temporal dynamics of specimen use

Herbarium-based publications dramatically increased over the past century (figure 2). However, given that the number of scientific publications overall has exponentially increased in recent decades as well (the "big literature phenomenon," Nunez-Mir et al. (2016)), we compared this trend to a background rate ("Plant Sciences" category in Web of Science). Herbarium-related publication rates have kept pace with the total plant science literature (test of difference between standardized publication rates from 2000 to 2017: $t = 0.268$, $df = 32$, $P = 0.79$).

Computational text analysis of 13,702 herbarium-related abstracts resulted in 22 meaningful topics, defined by an associated set of words with a high probability of co-occurrence (table 1; figure 3). Although topics varied in their prevalence, no single topic overwhelmingly dominated as a proportion of the overall corpus. Major topics in the herbarium-based studies were related to each other and clustered in meaningful ways (figure 4), with correlations between broader topic areas of plant morphology, taxonomy/systematics, floristics, history of botanists/collections, and biodiversity/global change. Groups of long-established

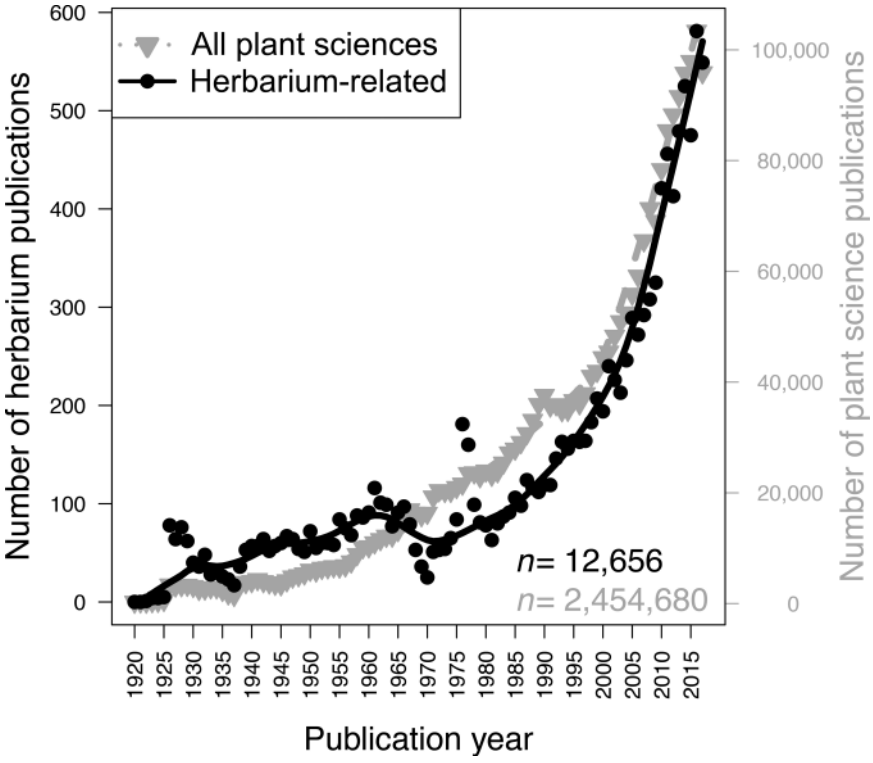

Figure 2. Number of articles using herbarium data published per year over the past century (1920–2017; Web of Science).

topics such as taxonomy-related (topics 1,6,11,12), floristics-related topics (topics 3,9,10,13), and morphology-related (topics 8,15,21,22) topics comprise 20%, 18%, and 12% of the corpus, respectively.

Despite only emerging in recent decades, studies involving new tools and approaches comprised 16% of all herbarium-based studies over the past 95 years: biodiversity informatics (e.g., niche models using digitized specimen occurrence data) 5% of corpus; global change biology (e.g., responses to past and future environments) 3% of corpus; DNA analyses (genetic data from specimens) 3% of corpus; and phytogeography and range dynamics (e.g., analyzing invasive species spread) 5% of corpus.

Topics represented in the herbarium-based literature were not evenly distributed over the past century, with marked declines in topics that once dominated the literature. For example, both history of botanists and collections (figure 5b) and taxonomic notes (figure 5d) have dropped in prevalence considerably. Although present in 15% of all studies across the 95-year period overall, their presence drops to only a few percent of published studies by 2016. Conversely, other topics have grown through time, and several topics emerged only in recent decades. As expected, the uses of herbarium specimens for DNA analyses materialized recently as a major topic. This topic was virtually absent from the literature prior to 1980 (figure 5d). Other notable emergent research areas are based in biodiversity and global change studies (figure 5f). Many research areas have maintained a relatively stable presence in the literature, including, for example, comparative morphology (figure 5c) and many taxonomically-related topics (figure 5d). Some topics follow more dynamic, non-linear trajectories, with some research areas clearly trending in particular decades (e.g., ploidy studies in the 1960s–1980s, figure 5c; species distributions, figure 5e).

Despite (or perhaps because of) the recent dramatic emergence of multiple new research areas, no research area clearly dominated in the past decade (figure 5a). Our analysis indicates a substantial increase in the number of topics that contribute non-trivially to the herbarium-based literature. Since 2000, there has been a functional diversification of the herbarium-based literature, with no topic present in more than 10% of the literature per year. The top five topics (ranked by prevalence in literature) accounted for nearly half (48%) of the literature during the earliest decade of our analysis (1920s), whereas the top five topics in the most recent decade (2010s) accounted for less than one-third (32%) of the literature.

Contextualizing herbarium use trends from automated content analysis

The topic model approach provided a holistic overview of the scope, major content areas, and temporal trends in herbarium-based research. The recent diversification of content areas (figure 5a) indicates the development of new approaches and techniques that enhance the longstanding primary uses of herbarium specimens, either by enabling deeper analyses of questions that were historically in the purview of herbarium-based research or innovations that generate new research questions altogether. The emergence of new topics in recent decades succinctly illustrates that herbarium specimens now serve new roles that were virtually absent or completely unimagined a century ago, such as isotopic analysis to determine changing atmospheric composition or DNA-based phylogeography. We provide evidence affirming the assertion that herbarium data remain relevant in modern plant biology research.

We argued in the introduction that one of the topic model's strengths is that it provides an unsupervised classification, that is, the researcher introduces no biases in *a priori* choice of topics to be analyzed. However, the approach may introduce its own biases. First, document selection depends on the researchers identifying a sufficient set of search strings to capture all the relevant

Table 1. ***Structural topic model results from 13,702 abstracts of herbarium-related studies. Topics are numbered in rank order by proportion of entire corpus (i.e., all abstracts analyzed) belonging to each topic, and including topic name, top 15 words with highest probability in each topic, and a general qualitative description of each topic. The dominant topic sets of high-probability words are generated by the algorithm while the general descriptions result from the authors' consensus of topics described in the abstracts associated with the topic (see supplemental materials).***

	Topic name	Top associated words	General description
1.	Taxonomic monographs and revisions	genus, taxonom, distribut, revis, base, key, descript, provid, present, includ, describ, field, illustr, discuss, recogn	Comprehensive treatments of taxon groups
2.	History of botanists and collections	botan, collect, garden, histori, natur, museum, work, centuri, univers, botanist, year, research, botani, public, first	Botanical history, focused on particular botanists, expeditions, or collections
3.	Species distributions	distribut, lichen, local, record, present, found, data, area, part, europ, rare, literatur, materi, itali, poland	Distribution of plant, lichen, and bryophyte taxa
4.	Collection summaries	collect, type, list, includ, given, number, materi, present, sheet, describ, nation, refer, famili, label, taxa	Summaries of single collections, especially those of particular botanists or type collections in a given natural history museum
5.	Biodiversity informatics and conservation	data, conserv, use, inform, collect, assess, databas, biodivers, research, can, provid, status, system, develop, need	Biodiversity databases; digitization of collections; statistical methods for biodiversity analysis; biodiversity conservation
6.	Typification/ Nomenclature	name, nomenclatur, type, lectotyp, design, materi, origin, publish, typif, synonym, valid, ident, describ, propos, lectotypif	Designation of type specimens and nomenclatural updates
7.	Phytogeography and range dynamics	distribut, area, habitat, rang, region, invas, forest, rich, divers, high, pattern, veget, climat, model, use	Species distributions in reference to geography, environmental gradients, and time (including non-native species spread)
8.	Morphometric studies	morpholog, charact, group, analysi, taxa, differ, variat, complex, distinct, use, within, variabl, separ, show, taxonom	Inter- and infra-specific studies based on statistical analyses of morphology (phenetics) (numerical taxonomy)
9.	Neotropical floristics	brazil, state, forest, famili, genera, collect, park, present, area, brazilian, rio, found, repres, distribut, record	Biodiversity studies in South America (especially Amazon region)
10.	Regional floristics and checklists	flora, taxa, endem, vascular, famili, florist, fern, record, region, genera, moss, checklist, includ, peninsula, list	Species lists and community descriptions at regional to local scales (preserves, parks, cities, physiographic regions, etc.)
11.	Taxonomic notes on genera/families	var, note, given, varieti, synonymi, genus, includ, addit, form, argentina, follow, discuss, literatur, descript, citat	Short reports and focused synopses of specific taxa (towards taxonomic monographs)
12.	New species descriptions	new, nov, india, describ, comb, combin, peru, colombia, costa, guinea, rubiacea, propos, ecuador, known, indian	Alpha taxonomy
13.	Local species observations and reports	north, america, state, mexico, american, counti, new, california, collect, nativ, unit, report, usa, mexican, known	Taxon occurrences at local or regional scales, primarily geopolitical units (especially county level in North America)
14.	Herbarium methodology and phytochemistry	use, sampl, method, dri, extract, materi, chemic, wood, result, differ, content, acid, preserv, contain, test	Techniques for specimen preservation; herbarium pest management; chemical properties of specimens
15.	Comparative morphology	leav, flower, fruit, leaf, differ, long, form, cell, infloresc, branch, stem, develop, shape, character, charact	Morphological studies within and between multiple taxa
16.	Global change biology	chang, flower, increas, time, seed, climat, year, temperatur, respons, phenolog, observ, differ, signific, use, period	Responses to past and future environmental change (especially atmospheric change: CO2, climate); phenological change through time; community- and population-level change
17.	Phytopathology	fungi, host, pathogen, isol, cultur, fungus, diseas, fungal, rust, caus, infect, parasit, associ, strain, collect	Studies on pathogens and their plant hosts
18.	Ethnobotany	use, medicin, tradit, local, famili, inform, knowledg, identifi, collect, part, district, survey, ethnobotan, peopl, interview	Traditional plant knowledge; economic and medicinal botany
19.	DNA analyses	sequenc, dna, phylogenet, molecular, use, clade, data, genet, analys, region, sampl, gene, within, support, lineag	Extraction, amplification, and analysis of DNA (especially molecular systematics)
20.	Algal floristics and taxonomy	island, coast, alga, collect, zealand, new, lake, record, mediterranean, pacif, marin, sea, water, coastal, archipelago	Marine biology (especially macrophytes, diatoms)
21.	Ploidy studies	popul, hybrid, chromosom, orchid, number, diploid, wild, cultiv, origin, genet, found, natur, cytolog, tetraploid, parent	Karyology, cytology, and palynology
22.	Morphology and anatomy	pollen, spore, structur, type, anatomi, grain, anatom, morpholog, studi, leaf, use, featur, section, electron, scan	Morphology and anatomy of specific structures, especially at micro-level

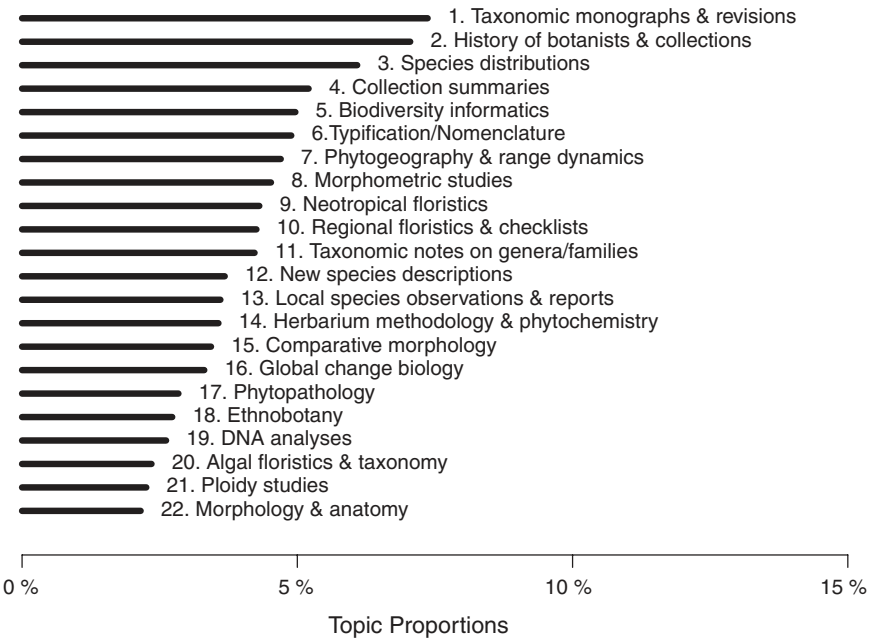

Figure 3. Topic proportions from structural topic model of herbarium-related literature (1920–2017). Topic proportions are the percentage of the total corpus that belongs to each topic. Topic names were defined from top words associated with each topic and holistic themes across the top abstracts related to each topic (see table 1).

documents. It may not ever be possible to include all relevant published literature. In our case, the search was based on the presence of the word "herbarium" and its derivatives. The automated selection of documents for inclusion in the analysis clearly does not detect all peer-reviewed published studies that made use of herbarium specimens and/or data. For example, some highly specialized journals may not be included in Web of Science or Scopus, and some may not have English language abstracts. Second, because only abstracts and titles are searched, some journals that specialize nearly entirely in herbarium-related research may not mention herbaria in the abstract or title; it is simply assumed that the authors made use of herbaria or perhaps indirectly stated only in the methods or acknowledgments. This may be especially true for relatively recent biogeographical studies that leverage big data from multiple biodiversity sources (combining observation- and specimen-based records) or data aggregators. In contrast, studies from disparate disciplines that make use of herbarium information may not be familiar with the conventional language in which the use of herbarium materials is normally described and are also missed in the automated search. Since the goal of topic modeling is not to enumerate all papers published, but to describe a best-fit set of topics within a given area of scholarly work, the key question is the extent to which the lack of the undetected papers in the literature corpus introduces bias in the importance of particular topics, or even prevents the identification of particular topics in the literature under consideration. There is not at present a clear method for addressing this question. Some degree of bias in articles selected for inclusion is very likely present in our current study. We think it quite unlikely that such bias would meaningfully alter the major results of this study. As topic modeling methods become more refined, greater confidence in the outcomes will be possible, and more nuanced questions will be addressable.

Because the automated topic model approach only provides a list of topics and related metrics, defined as sets of co-occurring words, the approach's value depends on discussion and interpretation of topic meaning. To complement this automated analysis, we therefore review these major research uses of herbarium specimens across a range of disciplines that emerged from our topic model, with particular attention to examples of novel recent use of specimens as sources for primary data.

Specimen-derived data in taxonomy and systematics

Taxonomy and systematics are the historic core areas of herbarium use, comprising roughly a fifth of the literature published over the last century. These uses have remained stable overall (except for "taxonomic notes"; figure 5d). Indeed, herbaria remain the major source for both the discovery and subsequent formal description of plant taxa new to science, with >50% of yet-to-be described species estimated to already be in herbaria awaiting description (Bebber et al. 2010).

New tools and approaches have revolutionized the study of taxonomic topics. Early studies in molecular biology and systematics pioneered methods and recognized the potential for extracting DNA from herbarium specimens (Rogers and Bendich 1985), but only recently have dried specimens been usable as primary sources of DNA on a massive scale (Buerki and Baker 2016, Bieker and Martin 2018). Recent developments in DNA barcoding (short DNA sequences to identify specimens) have the potential to revolutionize species identification and metagenomics and place new value on herbaria as huge repositories of genetic information to serve as sources for reference libraries (Kuzmina et al. 2017). The increased ability to analyze DNA contained in natural history collections has been aptly coined "museomics" (Buerki and Baker 2016).

Specimen-derived genotypic and genomic data

The emergence of museomics has enabled entirely new realms of research questions, using specimens to answer

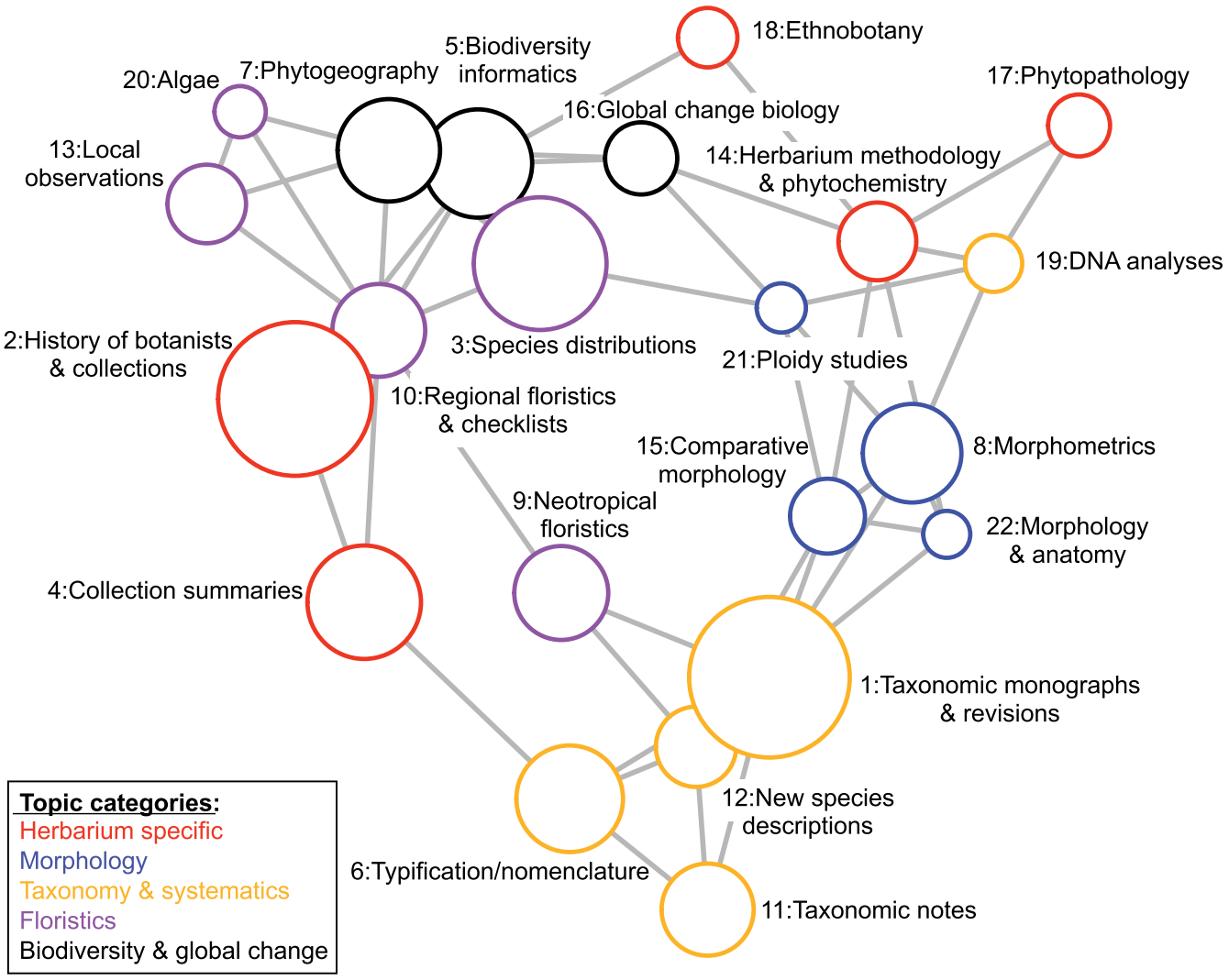

Figure 4. Topic correlation network showing associations between topics from the structural topic model of the herbarium-related literature (1920–2017). Topics near each other are more likely to appear together within abstracts. Node sizes are proportional to their topic proportions in the overall literature and are graphed using the Fruchterman-Reingold algorithm. Topics were grouped into color-coded categories as a visual aid.

questions at intraspecific and population levels. Sequencing of strains of the Irish potato famine pathogen from potato herbarium sheets is an early example that enabled a phylogeography of the pathogen in modern agriculture (Ristaino et al. 2001). Another pioneering study led to the discovery of native and non-native genotypes of the highly invasive weed, *Phragmites australis*, in North America (Saltonstall 2002). Since 2000, numerous other studies have leveraged historic specimens to answer ecological (Lavoie 2013) and evolutionary (Holmes et al. 2016) questions. Herbarium specimens have been recently shown to serve as material for "natural evolution experiments"; specimen genome sequences allowed quantification of mutation rates and their ecological consequences (Exposito-Alonso et al. 2018). Herbaria also hold potential as unplanned seed vaults for resurrection studies and the reintroduction of extinct wild and crop genotypic diversity (Leino and Edqvist 2010).

Specimen-derived occurrence data

The information provided on the labels of herbarium specimens is as important as the preserved organisms themselves (Merrill 1916). Digitization of natural history collections has dramatically transformed nearly all areas of collections-based research, but the effect of digitization of label data (including site localities) has been exceptionally profound. Paired with increasing availability of detailed environmental data across time and space, herbarium data have been the primary source for modeling species' distributions (Lavoie 2013). In particular, the introduction of maximum entropy (maxent) modeling has enabled accurate, fine-scale projections of species' potential distributions (Phillips et al. 2006).

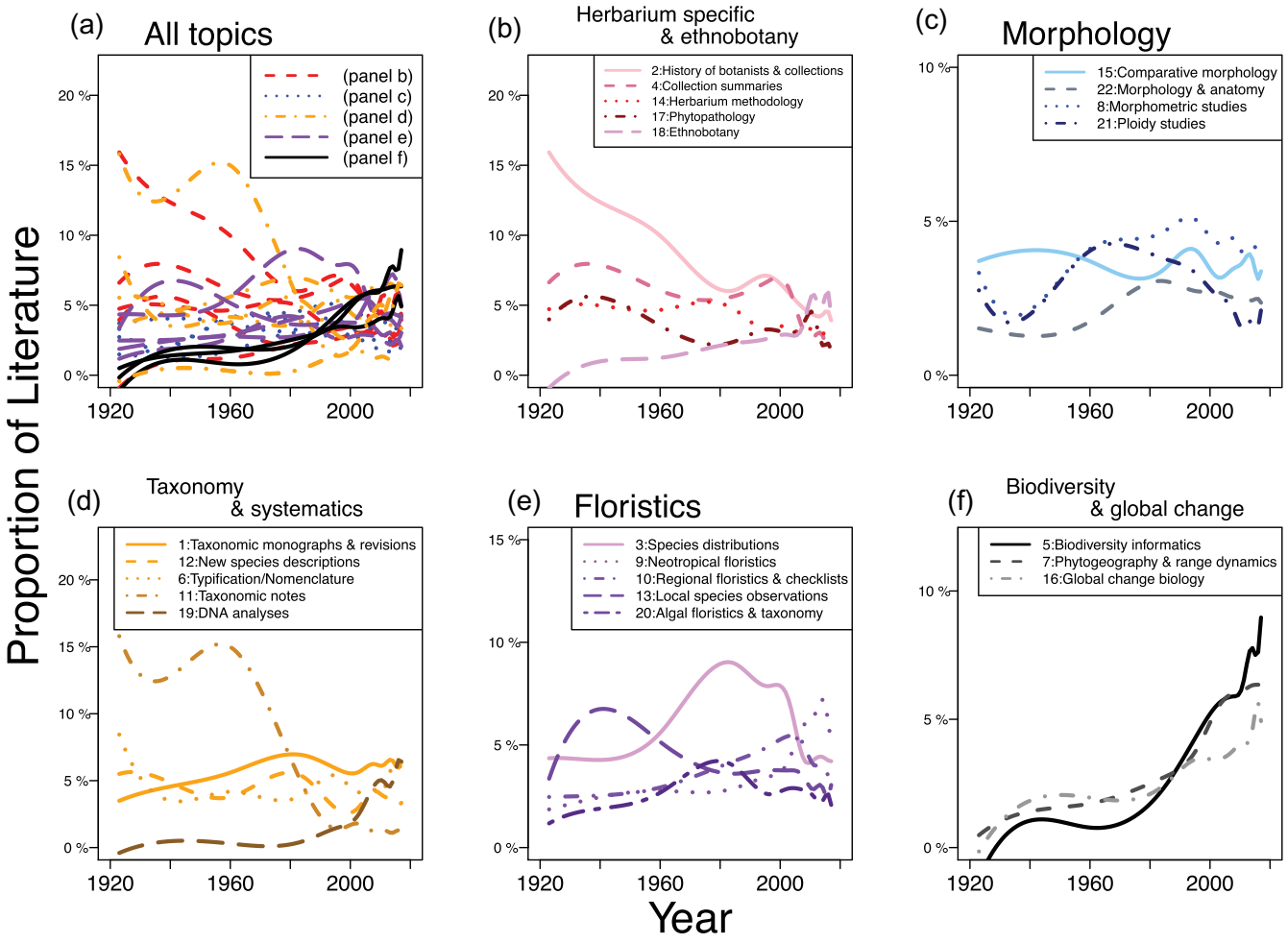

Figure 5. Topic prevalence over time in herbarium studies (1920–2017), as estimated from topic model of 13,702 abstracts. The vertical axis indicates the proportion of the literature in a given time period associated with a given topic. To aid in visual interpretation of temporal trends, topics were grouped into categories and color-coded (as in figure 4). Major temporal trends in the literature for (a) all 22 topics, (b) herbarium specific topics, phytopathology and ethnobotany, (c) topics based on morphological characters, (d) taxonomic topics, (e) floristics (studies on species in particular areas), and (f) topics related to biodiversity and environmental change.

As more georeferenced locality data go online (Nelson and Ellis 2018), it becomes increasingly powerful in ecological niche models for addressing core questions about the evolution and distribution of species and crucial environmental issues such as climate change, species invasions, and conservation (Soltis 2017).

Specimen-derived phenotypic data

Traditionally limited to species-level trait means for coarse taxonomic study, specimen phenotypes across time and space are increasingly being measured for finer-scale studies of evolutionary and ecological change. Influenced by developments in traits-based ecology and the advent of global change biology, herbarium-based studies on phenotypic change have surged, especially those measuring phenological responses to climate change (e.g., Willis et al. 2017). Coupling data across taxonomically diverse natural history collections can provide additional powerful approaches (eg., Kharouba and Vellend 2015). Specimen images enable greater accessibility, automation of functional trait measurements (Gehan and Kellogg 2017), and greater facility in taxonomic identification (Carranza-Rojas et al. 2017, Schuettpelz et al. 2017).

Extended phenotype data

At the frontier of unanticipated uses for herbarium specimens are attributes of the specimen that are not of the organism itself in the narrow sense. These studies include abiotic conditions when and where the specimen was collected, such as levels of industrial pollution in the soil (Rudin et al. 2017), changes in the atmospheric composition dating back to ancient Egypt (Beerling and Chaloner 1993)

and the Lewis and Clark expedition (Teece et al. 2002), and landscape level shifts in nutrient availability (McLauchlan et al. 2010). Other unanticipated studies focus on associated organisms, including plant pathogens (Ristaino et al. 2001) and invertebrates (Lees et al. 2011). Last, herbaria provide information on botanical history, world cultures, and historical uses of plants, with ethnobotanical specimens/artifacts as sources of otherwise unknown medicinal information (Souza and Hawkins 2017).

Relevance of specimens in a new era

Herbaria are critical resources for documenting biological and environmental change in the unfolding era of human domination of earth systems (Grinnell 1910, Pyke and Ehrlich 2010, Lavoie 2013). The general trends in plant collections are likely mirrored more broadly across all natural history collections (Suarez and Tsutsui 2004; Funk 2018, Schindel and Cook 2018). Winker (2004) argued that this shift towards biodiversity conservation and global change redefines the mission of natural history museums. Examples of research relating to urgent global change issues are often given as primary examples to assert relevance in order to justify continued support for natural history collections (Suarez and Tsutsui 2004). As uncovered by the topic analysis approach and reviewed in this discussion and elsewhere (Nualart et al. 2017), herbaria will become centers for conservation biology research and engage community participation and discussion (Ellwood et al. 2018). Mobilization of botanical information through virtual herbaria further enables direct conservation applications (Canteiro et al. 2019). Herbarium data will therefore not only provide insight into the past but also look to the future for restoration, stewardship, and environmental policy decisions.

We have also entered a new era with regard to digital access to collections (Drew et al. 2017). Many historic natural history collections were obtained through expeditions by European or American naturalists in past centuries, and therefore, increased data accessibility provides an ethically valuable repatriation of biodiversity data back to the communities from which these specimens were originally collected and emphasizes the value of open, shared scientific knowledge.

The role of continued collecting

Increasingly, herbarium-based studies rely upon large numbers of specimens collected across large spatial and temporal extents, but botanists decades ago realized the value of collecting many specimens per species (Woodson 1947). This conclusion argues strongly that existing collections, no matter how complete a representation of historical species diversity, will be far more valuable as research tools if collection is a continual activity. Despite this clear value, local plant collecting is on the decline (Prather et al. 2004a, Daru et al. 2018). Shifting research priorities also place less value on new, general collections (Prather et al. 2004b). Paradoxically, recent research relies increasingly heavily upon specimens that resulted from general collections. Recent calls have been made to re-evaluate specimen and data collection practices to maximize future use (Morrison et al. 2017), such as "holistic sampling," which involves comprehensively collecting specimens and data across multiple taxonomic groups in a community (Schindel and Cook 2018). To maximize their potential, specimens must be viewed in a broader framework, including, for example, traditional specimens, DNA vouchers, field notes, and images (the "extended specimen;" Webster 2017), to extract as much information as possible on the biology of the collected individuals through time and space. The use of new community science platforms for recording digital biodiversity data can powerfully complement traditional voucher collection (e.g., Heberling and Isaac 2018). Further, in an age of open and transparent data, herbarium vouchers are necessary, as specimens are primary data (Schilthuizen et al. 2015). Continued collecting clearly should be a priority.

Using natural history in predicting and managing nature's future

Ahead of his time, Parr (1939) stated that a museum "… can no longer claim justification by the mere existence of its collections". Natural history museums are at a turning point where they must redefine their institutional mission and relevance to society. Research in museums has shifted from a primary research mission of documenting the diversity of life to one of documenting biodiversity change (Winker 2004) and predicting the future of life. Museum scientists have begun adopting interdisciplinary roles in science advocacy, environmental justice, biodiversity conservation, and ethical discussions (Dorfman et al. 2018).

This review reports the increasingly diverse and novel uses of herbaria, and focuses on their potential in the future. Given the functional diversification of herbarium research we have documented, herbaria are well leveraged to encourage the cross-disciplinary synthesis necessary to advance research on urgent topics such as global change and restoration biology, while at the same time playing a central role in basic taxonomic, ecological, and evolutionary research. The increasingly diverse research value of herbaria documented here is concordant with a continuing general trend toward more integrative science, promising a strong role for herbaria in the future.

Acknowledgments

We thank the U.S. National Science Foundation (DBI-1612079), Roy A. Hunt Foundation, and MSU AgBioResearch for support. We are grateful to Scott Weingart and the Digital Humanities (Carnegie Mellon University Libraries) for advice on topic models. Thoughtful comments from the handling editor, Claude Lavoie, and three anonymous reviewers substantially improved earlier versions of this manuscript.

Supplemental material

Supplemental data are available at *BIOSCI* online. Additional data are available from the Dryad Digital Repository at *https://doi.org/10.5061/dryad.141jn7c*.

References cited

Bebber DP et al. 2010. Herbaria are a major frontier for species discovery. Proceedings of the National Academy of Sciences of the United States of America 107: 22169–22171.

Beerling DJ, Chaloner WG. 1993. Stomatal density responses of Egyptian *Olea europaea* L. leaves to CO_2 change since 1327 BC. Annals of Botany 71: 431–435.

Bieker VC, Martin MD. 2018. Implications and future prospects for evolutionary analyses of DNA in historical herbarium collections. Botany Letters 165: 409–418.

Blei DM, Ng AY, Jordan MI. 2003. Latent dirichlet allocation. Journal of Machine Learning Research 3: 993–1022.

Buerki S, Baker WJ. 2016. Collections-based research in the genomic era. Biological Journal of the Linnean Society 117: 5–10.

Canteiro C et al. 2019. Enhancement of conservation knowledge through increased access to botanical information. Conservation Biology 33: 523–533.

Carine MA, Cesar EA, Ellis L, Hunnex J, Paul AM, Prakash R, Rumsey FJ, Wajer J, Wilbraham J, Yesilyurt JC. 2018. Examining the spectra of herbarium uses and users. Botany Letters 165: 328–336.

Carranza-Rojas J, Goeau H, Bonnet P, Mata-Montero E, Joly A. 2017. Going deeper in the automated identification of herbarium specimens. BMC Evolutionary Biology 17: 181.

Cook JA et al. 2014. Natural history collections as emerging resources for innovative education. BioScience 64: 725–734.

Daru BH et al. 2018. Widespread sampling biases in herbaria revealed from large-scale digitization. New Phytologist 217: 939–955.

Dorfman E, Landim I, Kamei O. 2018. The future of natural history museums: general discussion. Pages 229–242 in E. Dorfman, (ed). *The Future of Natural History Museums*. Routledge, New York.

Drew JA, Moreau CS, Stiassny MLJ. 2017. Digitization of museum collections holds the potential to enhance researcher diversity. Nature Ecology and Evolution 1: 1789–1790.

Ellwood ER et al. 2018. Worldwide engagement for digitizing biocollections (WeDigBio): The biocollections community's citizen-science space on the calendar. BioScience 68: 112–124.

Exposito-Alonso M et al. 2018. The rate and potential relevance of new mutations in a colonizing plant lineage. PLOS Genetics 14: e1007155.

Farrell J. 2016. Corporate funding and ideological polarization about climate change. Proceedings of the National Academy of Sciences 113: 92–97.

Funk VA. 2018. Collections-based science in the 21st Century. Journal of Systematics and Evolution 56: 175–193.

Funk VA, Hoch PC, Prather LA, Wagner WL. 2005. The importance of vouchers. Taxon 54: 127–129.

Gehan MA, Kellogg EA. 2017. High-throughput phenotyping. American Journal of Botany 104: 505–508.

Grinnell J. 1910. The methods and uses of a research museum. Popular Science Monthly 77: 163–169.

Heberling JM, Isaac BL. 2017. Herbarium specimens as exaptations: New uses for old collections. American Journal of Botany 104: 963–965.

Heberling JM, Isaac BL. 2018. iNaturalist as a tool to expand the research value of museum specimens. Applications in Plant Sciences 6: e1193.

Holmes MW et al. 2016. Natural history collections as windows on evolutionary processes. Molecular Ecology 25: 864–881.

Kharouba HM, Vellend M. 2015. Flowering time of butterfly nectar food plants is more sensitive to temperature than the timing of butterfly adult flight. Journal of Animal Ecology 84: 1311–1321.

Kuzmina ML et al. 2017. Using herbarium-derived DNAs to assemble a large-scale DNA barcode library for the vascular plants of Canada. Applications in Plant Sciences 5: 1700079.

Lamstein A, Johnson BP. 2017. Choroplethr: Simplify the creation of choropleth maps in R. R package version 3.6.1. https://CRAN.R-project.org/package=choroplethr (accessed December 1, 2017).

Lang PLM, Willems FM, Scheepens JF, Burbano HA, Bossdorf O. 2019. Using herbaria to study global environmental change. New Phytol 221: 110–122.

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.

Lees DC, Lack HW, Rougerie R, Hernandez-Lopez A, Raus T, Avtzis ND, Augustin S, Lopez-Vaamonde C. 2011. Tracking origins of invasive herbivores through herbaria and archival DNA: the case of the horse-chestnut leaf miner. Frontiers in Ecology and the Environment 9: 322–328.

Leino MW, Edqvist J. 2010. Germination of 151-year old Acacia spp. seeds. Genetic Resources and Crop Evolution 57: 741–746.

McLauchlan KK, Ferguson CJ, Wilson IE, Ocheltree TW, Craine JM. 2010. Thirteen decades of foliar isotopes indicate declining nitrogen availability in central North American grasslands. New Phytologist 187: 1135–1145.

Meineke EK, Davies TJ, Daru BH, Davis CC. 2018a. Biological collections for understanding biodiversity in the Anthropocene. Philosophical Transactions of the Royal Society B 374: 20170386.

Meineke EK, Davis CC, Davies TJ. 2018b. The unrealized potential of herbaria in global change biology. Ecological Monographs 88: 505–525.

Merrill ED. 1916. On the utility of field labels in herbarium practise. Science 44: 664–670.

Monfils AK, Powers KE, Marshall CJ, Martine CT, Smith JF, Prather LA. 2017. Natural history collections: Teaching about biodiversity across time, space, and digital platforms. Southeastern Naturalist 16: 47–57.

Morrison SA, Sillett TS, Funk WC, Ghalambor CK, Rick TC. 2017. Equipping the 22nd-century historical ecologist. Trends in Ecology and Evolution 32: 578–588.

Nelson G, Ellis S. 2018. The history and impact of digitization and data mobilization on biodiversity research. Philosophical Transactions of the Royal Society B 374: 20170391.

Nualart N, Ibáñez N, Soriano I, López-Pujol J. 2017. Assessing the relevance of herbarium collections as tools for conservation biology. Botanical Review 83: 303–325.

Nunez-Mir GC, Iannone BV, Pijanowski BC, Kong N, Fei S. 2016. Automated content analysis: Addressing the big literature challenge in ecology and evolution. Methods in Ecology and Evolution 7: 1262–1272.

Parr AEE. 1939. On the functions of the natural history museum. Transactions of the New York Academy of Sciences 2: 44–58.

Phillips SJ, Anderson RP, Schapire RE. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190: 231–252.

Prather LA, Fuentes OA, Mayfield MH, Ferguson CJ. 2004a. The decline of plant collecting in the United States: a threat to the infrastructure of biodiversity studies. Systematic Botany 29: 15–28.

Prather LA, Fuentes OA, Mayfield MH, Ferguson CJ. 2004b. Implications of the decline in plant collecting for systematic and floristic research. Systematic Botany 29: 216–220.

Pyke GH, Ehrlich PR. 2010. Biological collections and ecological/environmental research: a review, some observations and a look to the future. Biological Reviews 85: 247–266.

R Core Team. 2017. R: A language and environment for statistical computing, version 3.4.1. Vienna.

Ristaino JB, Groves CT, Parra GR. 2001. PCR amplification of the Irish potato famine pathogen from historic specimens. Nature 411:695–697.

Roberts ME, Stewart BM, Tingley D. 2017. stm: R package for structural topic models. R package version 1.3.0. http://www.structuraltopicmodel.com.

Roberts ME, Stewart BM, Tingley D, Lucas C, Leder-Luis J, Gadarian SK, Albertson B, Rand DG. 2014. Structural topic models for open-ended survey responses. American Journal of Political Science 58: 1064–1082.

Rogers SO, Bendich AJ. 1985. Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. Plant Molecular Biology 5: 69–76.

Rudin SM, Murray DW, Whitfeld TJS. 2017. Retrospective analysis of heavy metal contamination in Rhode Island based on old and new herbarium specimens. Applications in Plant Sciences 5: 1600108.

Saltonstall K. 2002. Cryptic invasion by a non-native genotype of the common reed, Phragmites australis, into North America. Proceedings of the National Academy of Sciences of the United States of America 99: 2445–2449.

Schindel DE, Cook JA. 2018. The next generation of natural history collections. PLoS Biology 16: e2006125.

Schilthuizen M, Vairappan CS, Slade EM, Mann DJ, Miller JA. 2015. Specimens as primary data: museums and "open science." Trends in Ecology and Evolution 30: 237–238.

Schuettpelz E, Frandsen P, Dikow R, Brown A, Orli S, Peters M, Metallo A, Funk V, Dorr L. 2017. Applications of deep convolutional neural networks to digitized natural history collections. Biodiversity Data Journal 5: e21139.

Soltis PS. 2017. Digitization of herbaria enables novel research. American Journal of Botany 104: 1281–1284.

Souza ENF, Hawkins JA. 2017. Comparison of herbarium label data and published medicinal use: Herbaria as an underutilized source of ethnobotanical information. Economic Botany 71: 1–12.

Suarez A, Tsutsui N. 2004. The value of museum collections for research and society. BioScience 54: 66–74.

Teece MA, Fogel M, Tuross N, McCourt RM, Spamer EE. 2002. The Lewis and Clark Herbarium of the Academy of Natural Sciences, Part 3. Modern environmental applications of a historic nineteenth century botanical collection. Notulae Naturae 477: 1–16.

Thiers BM. 2018. The world's herbaria 2017: A summary report based on data from Index Herbariorum. Available from http://sweetgum.nybg.org/science/ih/ (accessed March 8, 2018).

Webster M. 2017. The Extended Specimen. Pages 1–9 in M. Webster, (ed). *The Extended Specimen: Emerging Frontiers in Collections-Based Ornithological Research*. CRC Press, Boca Raton.

Willis CG et al. 2017. Old plants, new tricks: Phenological research using herbarium specimens. Trends in Ecology and Evolution 32: 531–546.

Winker K. 2004. Natural history museums in a postbiodiversity era. BioScience 54: 455–459.

Woodson RE. 1947. Some dynamics of leaf variation in Asclepias tuberosa. Annals of the Missouri Botanical Garden 34: 353–432.

J. Mason Heberling (heberlingm@carnegiemnh.org) is an Assistant Curator of Botany, and Stephen J. Tonsor is the Director of Science & Research at the Carnegie Museum of Natural History. L. Alan Prather is the Director of the MSU Herbarium and Associate Professor in the Department of Plant Biology at Michigan State University. JMH is a plant ecologist whose research is focused on the ecophysiology of native and invasive plants in eastern North American forests. SJT is an evolutionary ecologist who considers adaptation to changing environments through genetics, physiology and life history allocation. LAP is a plant systematist whose research combines field studies, molecular data, and morphological traits to study phylogenetics and phenotypic diversity.