Applications in Plant Sciences

REVIEW ARTICLE

INVITED SPECIAL ARTICLE
For the Special Issue: Green Digitization: Online Botanical Collections Data Answering Real-World Questions

Herbarium data: Global biodiversity and societal botanical needs for novel research

Shelley A. James[1,7], Pamela S. Soltis[2], Lee Belbin[3], Arthur D. Chapman[4], Gil Nelson[5], Deborah L. Paul[5], and Matthew Collins[6]

Manuscript received 3 October 2017; revision accepted 30 December 2017.

[1] National Herbarium of New South Wales, Royal Botanic Gardens and Domain Trust, Mrs Macquaries Road, Sydney, New South Wales 2000, Australia

[2] Florida Museum of Natural History, University of Florida, Gainesville, Florida 32611, USA

[3] Atlas of Living Australia, CSIRO, Clunies Ross Street, Acton, Australia Capital Territory 2601, Australia

[4] Australian Biodiversity Information Services, Ballan, Victoria 3342, Australia

[5] iDigBio, Florida State University, Tallahassee, Florida 32306, USA

[6] Advanced Computing and Information Systems, University of Florida, Gainesville, Florida 32611, USA

[7] Author for correspondence: shelley.james@rbgsyd.nsw.gov.au

Citation: James, S. A., P. S. Soltis, L. Belbin, A. D. Chapman, G. Nelson, D. L. Paul, and M. Collins. 2018. Herbarium data: Global biodiversity and societal botanical needs for novel research. *Applications in Plant Sciences* 6(2): e1024.

doi:10.1002/aps3.1024

Building on centuries of research based on herbarium specimens gathered through time and around the globe, a new era of discovery, synthesis, and prediction using digitized collections data has begun. This paper provides an overview of how aggregated, open access botanical and associated biological, environmental, and ecological data sets, from genes to the ecosystem, can be used to document the impacts of global change on communities, organisms, and society; predict future impacts; and help to drive the remediation of change. Advocacy for botanical collections and their expansion is needed, including ongoing digitization and online publishing. The addition of non-traditional digitized data fields, user annotation capability, and born-digital field data collection enables the rapid access of rich, digitally available data sets for research, education, informed decision-making, and other scholarly and creative activities. Researchers are receiving enormous benefits from data aggregators including the Global Biodiversity Information Facility (GBIF), Integrated Digitized Biocollections (iDigBio), the Atlas of Living Australia (ALA), and the Biodiversity Heritage Library (BHL), but effective collaboration around data infrastructures is needed when working with large and disparate data sets. Tools for data discovery, visualization, analysis, and skills training are increasingly important for inspiring novel research that improves the intrinsic value of physical and digital botanical collections.

KEY WORDS biodiversity data; biodiversity standards; global change; herbarium collections; informatics.

Botanical collecting during the past 450 years has resulted in the permanent housing of more than 381 million specimens in over 3000 herbaria scattered across the globe (Krishtalka et al., 2016; Thiers, 2017). Such collections continue to be gathered and are a valuable record of biodiversity across biomes and through time, from deep time to yesterday (Gardner et al., 2014; Page et al., 2015; Holmes et al., 2016; Willis et al., 2017). The increasing challenges of funding highlight the need for herbaria and botanical collections to improve curation efficiency. With ever-improving digital technologies enhancing database and image-capture workflow efficiencies, the complete cataloging of collections is now possible. Collections data can be published and shared with data aggregators, allowing for global discovery, synthesis, and prediction. This paper provides an overview of how aggregated, open access botanical and associated genetic, trait, environmental, and ecological data sets are available for assessing the impacts of global change on communities, organisms, and society; predicting future impacts; and helping to drive the remediation of change. The challenges of using such data are discussed.

RESEARCH USE OF HERBARIUM DATA

Herbarium specimens and their data are, for the most part, verifiable, repeatable, sustainable, and persistent (Page et al., 2015; Holmes et al., 2016). Temporal data across taxonomic groups, communities, and habitats enable assessment of changes in species distributions, dispersal ability, or clade differences. Interactions within

and between taxa can be interpreted, providing information about species associations and community assemblages through space and time. Historical and reliable baseline data from collections are needed to build robust predictive models for various taxon-level or functional-group global change responses (e.g., Willis et al., 2017). Herbarium collections and the data they hold are valuable for more traditional studies of taxonomy and systematics, but also for ecology, bioengineering, conservation, food security, and the human social and cultural elements of scientific collection (Culley, 2013; Heberling and Isaac, 2017; Soltis, 2017; Willis et al., 2017). Botanical specimens provide baseline data for basic to applied research applications (Appendix S1; Chapman, 2005). Effects of global change on human health and ecosystem services can be studied using primary biodiversity data, with topics such as the distribution and spread of disease vectors, flora and fauna of economic importance, and the introduction, impact, and spread of non-native and invasive species (Arnaud et al., 2016; McGeoch et al., 2016).

Significant and irreparable changes to Earth's ecosystems due to global change can be seen by examining the shifts in species distributions and community structure in space and time (IPCC, 2014). By incorporating and combining data sources for environmental factors with biological data, primary biodiversity data from herbaria, and other natural history collections, along with other informative data such as tree ring data, observational records, and phenological and other trait data, analyses can be performed to gain an improved understanding of the impacts of change on global biodiversity. Such research is increasingly requiring collaborative, interdisciplinary science (AIBS, 2015a; Soranno et al., 2015). Botanical data can be used as training data for developing statistical models to predict the way changes will affect organisms. Such models may be used as conservation and policy tools to lessen or mitigate the effects of global change on biodiversity and food security (Jarvis et al., 2008; Guisan et al., 2013). To improve model performance, data gap analysis and focused digitization efforts for particular geographic regions or taxonomic groups may be needed to ascertain data completeness for baseline species distribution assessments (Pino-Del-Carpio et al., 2014).

Paleobotanical and paleoecological data, including fossil pollen, stomate size, and evidence of leaf damage by herbivores, can be used to explore species and ecological assemblages over time and against changing environmental parameters (e.g., Strömberg et al., 2013; Kohn et al., 2015; Maguire et al., 2016). Integration of neontological and paleontological biodiversity data, linking with literature-based occurrence data found in resources such as the Paleobiology Database (PaleoDB; https://paleobiodb.org/) and using application programming interfaces (APIs) and other cyberinfrastructure services such as those becoming available through the enhancing Paleontological and Neontological Data Discovery API project (ePANDDA; https://epandda.org/), is helping to answer deep-time to present-day global change research questions. The ability to study communities of organisms through time will require continued coordination of the development of digitization workflows and best practices between collections of different taxonomic groups within both neontological and paleontological collections, with data standardized for efficient integration, aggregation, and downstream use in analyses.

Primary biodiversity data can be used to study changes in communities, temporally and spatially, and shifts in community associations within and between taxonomic groups (Morueta-Holme et al., 2016). As the volume of biodiversity data from multiple collections of taxonomic and geographic breadth is aggregated, along with supporting observational data records, an assessment of global changes in biological community organization and structure is enabled. Resurveys of biodiversity and the pooling of data across geographic regions, in comparison with legacy data, can be used to assess long-term shifts in community structure (Verheyen et al., 2017). Long-term monitoring projects (e.g., the Long Term Ecological Research Network projects [LTER; https://lternet.edu/] and National Ecological Observatory Network [NEON; http://www.neonscience.org/] stations in the United States, the Terrestrial Ecosystem Research Network [TERN; http://www.tern.org.au/] in Australia, and the Chinese Ecological Research Network [CERN http://cnern.cern.ac.cn/en]) are providing open access to long-term standardized ecological and biological data sets, with the historical data within herbarium collections providing the historical baseline. Phenological data associated with biological collections document changes in seasonality over time and provide insight as to the effect on community associations in a broader context (Davis et al., 2015; Willis et al., 2017).

Species and community assemblages can be indicators of habitat health. Changes in community composition across space and time may be correlated with the appearance of invasive species, changes in environment, or human activity. Baseline documentation of communities as found within botanical collections data sets can be used for restoration or rehabilitation purposes and may be useful for determining surrogate taxa (Weirauch et al., 2017). Collections data sets consisting of both paleological and neontological specimen data are increasingly essential for conservation purposes (e.g., Ponder et al., 2001; Pino-Del-Carpio et al., 2014; Barnosky et al., 2017). Organizations such as the International Union for Conservation of Nature (IUCN), World Wide Fund for Nature (also known as the World Wildlife Fund; WWF), The Nature Conservancy, NatureServe, and others benefit from biological collections data and are primarily interested in habitat and species evaluations. Primary biodiversity data are critical for species conservation assessments such as the IUCN Red List (Brummitt et al., 2015) and delineation of protected areas. Biological collections data can be used to provide data for proactive systematic conservation planning or for rehabilitation or restoration efforts, such as the delineation of climate refugia, buffer zones, and corridors. "Alpha" (location of hotspots, design of reserves, restoration assessment) or "beta" (specific species protection, reintroduction programs) conservation questions and policy development can be determined using herbarium voucher specimen data (Soberon et al., 2000). Niche or species distribution modeling using biological collections data can assist with anticipating taxon range shifts, future needs, and restoration parameters due to changes in climatic regimes (Guisan et al., 2013). An example is the Australian-based Restore and Renew project (https://www.rbgsyd.nsw.gov.au/science/restore-renew), which relies on herbarium records to plan fieldwork for gathering voucher and tissue collections across the entire geographical and ecological distribution of the study species. Surrogate species distributions can also be used to assess rare and endangered species distributions such as the historical and current distribution of communities and host taxa (Morales-Castilla, 2015; Weirauch et al., 2017). The early detection of incipient invasive species and documentation of the movement and initial invasion point of invasive species depend on primary biodiversity data. Species distribution models can also be used to better understand biological invasions (Guisan et al., 2013) and to identify potential biological control agents (Sutherst, 2014). However, the use of species distribution

modeling as a tool for successful conservation planning and policy is often limited by data quality, data availability, and data bias (Cayuela et al., 2009; Elith and Leathwick, 2009). The fitness for use of primary biodiversity data for species distribution modeling (Anderson et al., 2016), agrobiodiversity (Arnaud et al., 2016), and alien and invasive species (McGeoch et al., 2016) has recently been reviewed by Global Biodiversity Information Facility (GBIF) Task Groups. Linkages between specimen collections and conservation information about taxa can be useful for researchers, land managers, policy-makers, and others interested in protected species or areas. For example, linking specimens of taxa with information about IUCN Red List status, federal or state endangered species listings, or Convention on International Trade in Endangered Species (CITES) restrictions supports research and education.

Linking collections data to phylogenetic data enables the assessment of how global change has influenced or may influence genetic and/or phylogenetic diversity of communities spatially and temporally (Holmes et al., 2016; Soltis, 2017; Allen et al., unpublished manuscript). By linking collections data to different landscape features and assessing how young versus old lineages diversified across space and time, evolutionary trajectories of clades can be analyzed. Biodiversity hotspot analysis can be used to determine regions of interest for further exploration or protection, as well as to supplement the testing of diversity hypotheses and biogeographic theories (e.g., Phillips et al., 2011). Hypotheses of community homogenization, both taxonomic and phylogenetic and including paleontological or pre-industrial versus modern communities, can be tested.

Herbarium data fitness for use

Primary biodiversity data, including herbarium data, are not always research-ready, and the fitness for use of data will depend on the requirements of each research project and the availability and accessibility of information within herbarium collections. Biodiversity data have been described as biased, fuzzy, haphazard, unstandardized, non-random, incomplete, and unique because of collecting bias and/or digitization gaps, and subsequently require quality assessment (Soberon et al., 2000, 2007; Hortal et al., 2007; Meyer et al., 2015; Gueta and Carmel, 2016; Willis et al., 2017; Daru et al., 2018). Predictive modeling or other statistical analyses may help fill such gaps (Hortal et al., 2007; Chao et al., 2009), but further sampling and digitization efforts are still needed to address spatial, temporal, taxonomic, and data quality gaps and shortcomings (Berendsohn and Seltmann, 2010; Ariño et al., 2016; Troudet et al., 2017).

Taxonomic limitations that users of biodiversity data need to be aware of include the following:

1. Taxonomic or nomenclatural expertise is underestimated for data interpretation (Soberon et al., 2000). Issues associated with taxonomic revision and interpretation are variable across taxonomic groups (Hortal et al., 2007; Troudet et al., 2017).
2. Data aggregators conform to a single synthetic management classification, or taxonomic authority file, for indexing, searching, and discoverability (e.g., the GBIF Backbone Taxonomy), which may result in bias if the raw data are not also considered by researchers (Murray et al., 2017). Lags associated with taxonomic consensus and curation of both physical specimens and data also delay data and updates shared through biodiversity data aggregators (e.g., Bebber et al., 2010).
3. There is often a selective focus on taxonomic groups within collections resulting from the specialty of curators and funded projects (e.g., Daru et al., 2018).
4. Collections often have a selective focus based on taxonomic uniqueness (e.g., endemics), rarity, or economic value. Rare species can often be overrepresented in collections and databases, but are increasingly less represented in more modern collections due to permitting issues. Common species and/or introduced species, by comparison, are often not collected due to the limited capacity and storage of institutions despite the importance of recording current-day phenological, morphological/anatomical, environmental, or distributional outliers.

Spatial limitations can include the following:

1. Collecting effort is not randomly or regularly distributed (Soberon et al., 2000; Davis et al., 2015), and biodiversity patterns are scale-dependent and sensitive to spatial resolution (Soberon et al., 2007).
2. Not all specimens have adequate associated locality information, including named locations with distance and direction data, uncertainty information associated with the location, or metadata such as the geodetic datum used to determine latitude and longitude (Chapman and Wieczorek, 2006).
3. Data generalization of sensitive taxa or taxa from sensitive locations can lead to errors in analysis if not documented and detected by users (Chapman and Grafton, 2008).
4. Collections are often focused on nearby accessible geographic regions, such as the proximity to research institution or along roads (Prendergast et al., 1993; Davis et al., 2015; Meyer et al., 2015; Daru et al., 2018), and can be limited in scope due to the regional mission bias of institutions. Intensive localized collection may result from expedition events, ecological assessment, permanent plots or long-term monitoring, or hotspot analysis (e.g., parks, mountain ranges, wetlands).
5. Geographic collection limitations can be associated with historical, political, funding, or social barriers (e.g., Crawford and Hoagland, 2009).
6. Downscaling information within biodiversity data sets when georeferencing accuracy is low may be problematic for localized analyses.

Temporal limitations can include:

1. Collecting effort is not evenly distributed in time (Soberon et al., 2000; Davis et al., 2015) and has declined since the mid-20th century (Gardner et al., 2014).
2. Collectors have been, and continue to be, biased in activity during certain time periods both on annual and historical time scales (Prendergast et al., 1993; Daru et al, 2018).
3. Due to the interests of collectors, institutions, and funding agencies, taxa and geographic regions will have temporal biases within collections.
4. Dates may have imprecise month/season/year time ranges, and social and societal differences of date recording (e.g., day/month/year versus month/day/year) may cause confusion in transcription and reduce data fitness for use.

Absence data are important for statistical modeling algorithms. However, such data are rare or not easily discoverable within herbarium data sets. Lack of collections of a taxon at a place and time cannot imply absence. Natural heritage programs and collectors often do capture observational absence information when they search for a taxon, but that information may not be mobilized and is, therefore, effectively unavailable. This is largely a limitation resulting from a data provider's inability to mobilize absence data in standardized fashion, and of the reliability of absence data. The Darwin Core Standard (DwC; http://rs.tdwg.org/dwc/index.htm), the primary international standard for encoding and exchanging specimen and observation data (Wieczorek et al., 2012), has the term 'occurrenceStatus' that accepts values of 'present' and 'absent,' but this standard postdates most collections records. There are DwC terms for sampling effort ('samplingEffort'), sampling methodology ('samplingProtocol'), and measurement or fact concept ('measurementOrFact'), for example, but data are often recorded inconsistently in a comments or remarks text field. Community efforts to develop controlled vocabularies will eventually help to address this issue.

TDWG data quality standards

Data trust and reliability, even for voucher specimen data, must be evaluated for fitness for use in each research use case (Ariño et al., 2016). There are several community approaches to assess herbarium specimen data quality and biodiversity data in general (e.g., Robertson et al., 2016; Morris et al., 2017). The Biodiversity Information Standards (TDWG) Data Quality Interest Group (DQIG; https://github.com/tdwg/bdq) was established in 2014 with the goals of assessing data quality and assisting with the standardization of the data delivered by aggregators and others. The aims of the DQIG are to establish a framework for "data quality" (Veiga et al., 2017); to standardize how specimen (and observation) records can be evaluated, amended, and reported; and to develop a set of profiles for use cases such as for species distribution modeling. This work focuses on the critical data dimensions of name, space, and time. The work of the DQIG has highlighted the problems associated with the lack of controlled vocabularies within the Darwin Core (DwC) standard. Unconstrained values used among the DwC terms mean that more tests are necessary to detect problems, some data problems cannot be detected, and scientists find it difficult to evaluate the data prior to research use. For example, while five values for dwc:basisOfRecord have been suggested in the standard (e.g., 'preservedSpecimen,' 'humanObservation'), GBIF had (as of June 2017) 2483 distinct values for that term. The outcome of the DQIG's activities will be the development of standard tools for herbarium and natural history collections and record data sets to enable improvement of data quality and fitness for use for a wide range of research questions.

Standard tests being developed by DQIG will be implemented by data collectors for use in the field; by data aggregators such as Integrated Digitized Biocollections (iDigBio; https://www.idigbio.org/), the Atlas of Living Australia (ALA; https://www.ala.org.au), and GBIF (https://www.gbif.org/); by ancillary services such as Kurator (Morris et al., 2017; http://wiki.datakurator.org/); by data users; and by herbarium data custodians. This will provide concise and consistent information for biodiversity data evaluation for different research and data use needs. Most data aggregators currently use test algorithms that report on various potential issues associated with data records, but each aggregator has its own suite of algorithms and reporting methods. Standardization of the tests and resulting assertions, how they are reported, and how these reports as annotations travel with the records are fundamental requirements for efficient research and area management.

Streamlining field data collection

As technology advances and digital tools are increasingly robust under field conditions, the capture of data in electronic rather than analog format is more efficient and accurate. Such born-digital data are critical for avoiding further backlog of data transcription in herbarium collections and for efficient downstream incorporation into collection management systems and data aggregators. Digital data capture leads to improved workflows, avoiding errors in transcription and enabling data to be available in a timely manner for global and societal scientific use. Digital technologies and mobile app development allow for locality data (Global Positioning System [GPS]), field images, and other field data elements to be automatically captured and linked, including standardized picklists and vocabularies (e.g., BioCollect; https://www.ala.org.au/biocollect/). One recently developed resource now in use is the Biocode Field Information Management System (FIMS; http://www.biscicol.org). Using this online tool, researchers can develop and customize their data collection protocol, select field headers (terms) from current data standards (e.g., DwC), and then output selected terms in a ready-to-use spreadsheet. This system includes the definitions for the terms as well as the data types (e.g., date, text, numeric) expected for each field. In addition, FIMS assigns globally unique identifiers (GUIDs) to each record in the generated template. The FIMS system incorporates several data standards, making it easy for researchers to integrate data. Once the spreadsheets are completed, data can be uploaded via the FIMS validation tool to check data quality and adherence to the expected data standards. The use of QR codes or barcodes with an embedded, computer-readable GUID equating to a DwC field, such as 'eventID,' along with a human-readable collection number attached to each element of a collection—from the field notebook to the collection tags, to tissues for molecular analysis and images of specimens—assists in the automated linkage and sharing of collections data.

Biodiversity data are collected with a particular use in mind. Additional information beyond the initial application will almost certainly support far broader applications into the future. It is important to recognize that the future of the data can never be fully anticipated, so the collection of additional data and metadata in the field is always a wise investment (Morrison et al., 2017). Even historically, J. Grinnell commented on how the value of collections and the data therein may not be realized in the immediate future (Grinnell, 1910), so he developed and implemented a detailed protocol for recording field observations (Grinnell, 1912).

Streamlining analysis of aggregated data

The size and scope of aggregated digital data have exploded and will continue to grow with efforts to digitize collections and collect digital biological data directly in the field. Combining large, diverse data sets is currently challenging due to limitations in standards and lack of consistent vocabularies and metadata between research fields. The development of ontologies (e.g., Walls et al., 2014) will help with reducing such barriers between biodiversity

resources. However, traditional analysis tools (e.g., spreadsheets, laptops, and databases) have struggled to manipulate the millions of records some research questions require. An approach to addressing this need is to build biodiversity data infrastructures for analyses and not just data aggregation (Poelen et al., 2014). One example is Global Unified Open Data Access (GUODA; http://guoda.bio/), a collaboration between developers and technical staff at Encyclopedia of Life (EOL; http://eol.org), iDigBio, and freelance software engineer Jorrit Poelen. An infrastructure based on Apache Spark (Zaharia et al., 2016) and biodiversity data sets such as EOL, iDigBio, GBIF, and the Biodiversity Heritage Library (BHL) is available for application developers and data analysts to build tools and services providing whole biodiversity data set analytics to explore broad biodiversity questions. As two proofs of concept, EOL and Poelen have developed Fresh Data (http://gimmefreshdata.github.io), a tool to discover and follow records in biodiversity archives that match specific geospatial, temporal, taxonomic, and trait constraints, and Effechecka (http://www.effechecka.org/), in which taxonomic checklists and occurrence lists are returned.

Combining analysis with aggregation of data allows for pattern searching, such as duplicate record resolution, outlier detection of specimen data (e.g., collection outside of collector or environmental range), and batch georeferencing. The work being done by the TDWG DQIG in collaboration with data aggregators to standardize basic data description, output, and data transfer will assist with streamlining such applications.

Digitized images of herbarium specimens for data analysis are increasingly available through biodiversity data aggregators. Aggregated herbarium image data are being utilized for projects such as the automated identification of herbarium specimens (e.g., Carranza-Rojas et al., 2017; Schuettpelz et al., 2017) and phenological studies (Willis et al., 2017). Historical images available through BHL are becoming increasingly linked to other data sources such as the EOL, ALA, GBIF nodes using the ALA platform, and other sites through community tagging on Flickr (https://www.flickr.com/people/biodivlibrary/). An example is the tagging of images with locality and taxonomic information from *Curtis's Botanical Magazine* (https://www.flickr.com/photos/biodivlibrary/collections/72157681766674633/) for linkage with the herbarium specimens found in ALA.

DISCUSSION

Use of research to drive digitization efforts

The Thematic Collections Networks (TCNs) funded through the U.S. National Science Foundation's Advancing Digitization of Biodiversity Collections (ADBC) program provide examples of compelling biodiversity hypotheses to be tested through funded digitization efforts. Novel research hypotheses, geographic and taxonomic themes, and societal demands of health and human services are needed to motivate future digitization and funding, and drive sustainability of collections digitization. GBIF established a task group to address the need to discover biocollections data not yet mobilized (Krishtalka et al., 2016), and others have proposed recommendations to the community and data aggregators for bridging biodiversity data gaps (Berents et al., 2010; Faith et al., 2013; Ariño et al., 2016; Geijzendorffer et al., 2016). Ultimately, the biodiversity data community needs to ask how herbaria, curators and researchers, and policy-makers should be playing a larger role in driving digitization efforts, and whether recognized data gaps should be preferentially addressed regardless of current research priorities. Improved access to biodiversity portal search data statistics or loan and collection use requests may help support digitization efforts. Including digitization as a component of museum or herbarium accreditation processes (e.g., American Alliance of Museums, National Standards for Australian Museums and Galleries) and strategic planning may help to drive systematic digitization, quality control, and the inventory of botanical collections. This includes encouraging botanical collections worldwide to provide and update information about their institution, holdings, and taxonomic expertise in online resources such as Index Herbariorum (http://sweetgum.nybg.org/science/ih/), the Global Registry of Biodiversity Repositories (GRBio; http://grbio.org/), and iDigBio's U.S. Collections list (https://www.idigbio.org/portal/collections). Ensuring that newly collected data are discoverable and fit for broad reuse requires the community to foster, adopt, and update collection and data gathering best practices and standards through the activities of organizations such as TDWG.

Education and training needs

Researchers, in particular those early in their careers, need greater exposure to the value of herbarium and biodiversity data available through collections and biological data aggregators. Researchers also need to build skills to be able to interrogate and utilize the available data. Hampton et al. (2017) recently outlined five capstone skills needed by environmental scientists, and by extension, biodiversity scientists and data curators: data management and processing for reproducibility, analysis, software skills, visualization, and communication methods for collaboration and dissemination. An awareness and understanding of the biases, issues, and limitations of the data that are provided are critical for appropriate use of biodiversity data (Gueta and Carmel, 2016; Hampton et al., 2017). Such data literacy and data evaluation skills are needed in the community, from the undergraduate to professional level, for the analysis of large, combined biodiversity data sets (AIBS, 2015b; Hampton et al., 2017). Efforts are underway with the Biodiversity Literacy in Undergraduate Education (BLUE; http://biodiversityliteracy.com) Network, which is developing curricula and building a community network to develop data literacy standards for future research career professionals and the public, who need to be able to interpret the results from global scientific research using botanical and natural history collections data. Reproducible science, appropriate citation, and open data should be priorities in training efforts of the biodiversity community (Bishop and Hank, 2016), with the FAIR Guiding Principles (i.e., Findability, Accessibility, Interoperability, and Reusability) for scientific data management and stewardship as a guiding infrastructure (Wilkinson et al., 2016).

Born-digital data and analysis

With technological advances (August et al., 2015), more efficient methods are being developed and shared for the collection of new data (e.g., iDigBio-sponsored Field to Database [https://www.idigbio.org/wiki/index.php/Field_to_Database] and Georeferencing for Research Use [https://www.idigbio.org/wiki/index.php/Georeferencing_for_Research_Use] workshops). The lessons learnt and the gaps discovered because of the digitization of existing

collections need to be carried into future collecting efforts and expeditions. Data curation profiling of biocollections may assist managers and researchers to capture information that informs data curation beyond the technical needs for data ingestion (Bishop and Hank, 2016). Often, collection managers are one step removed from scientists and citizen scientists who have collected or are collecting the specimens. A data curation profile "captures requirements for specific data generated by researchers articulated by the researchers themselves" (Bishop and Hank, 2016), providing metadata that can aid in linking otherwise disparate data sets and making broader reuse of valuable data. Engagement and training of non-collections personnel, such as environmental scientists and ecologists, is increasingly important for specimen collection and biodiversity data capture (Ward et al., 2015). DNA sequence capture and molecular ecology alone will not resolve the understanding of biodiversity (Creer et al., 2016).

Internationalization: Engaging and enhancing global digitization

Many large herbaria in the developed world are developing and implementing digitization goals, often with a mandate driven by institutional needs. However, much of the regional diversity and often taxon-specific collections highly valuable for research are found in smaller local herbaria and museums, which may lack the infrastructure and resources needed to digitize collections (Casas-Marce et al., 2012). This disparity results in major gaps in primary biodiversity data sets. A critical community goal is to incentivize more institutions to mobilize collections data for physical specimen inventory, curation, and research; this is being done primarily through funding, as well as through training and infrastructural support (Canhos et al., 2015). Biodiversity aggregators are increasingly interested in the use of collections data in research to drive their sustainability, and the physical herbarium collections need data-use metrics about their collections to maintain funding and institutional support for the continued digitization and publishing of data they transcribe and curate. Appropriate acknowledgment of herbarium collections and their data sets in publications (Rouhan et al., 2017), the use of object and data record identifiers for data tracking (James, 2017), and the development of community standards for citation (e.g., working groups of the Research Data Alliance [Rauber et al., 2015], TDWG Natural Collections Descriptions Interest Group [http://www.tdwg.org/activities/ncd/]) will help to enhance the sustainability of digitization and botanical data mobilization into the future. Appropriate attribution shows advocacy for the continued preservation, expansion, and availability of the physical and digital botanical collections curated by herbaria into the future (Suarez and Tsutsui, 2004; Winston, 2007).

ACKNOWLEDGMENTS

Integrated Digitized Biocollections (iDigBio) is funded by grants from the U.S. National Science Foundation's Advancing Digitization of Biodiversity Collections program (Co-operative Agreements EF-1115210 and DBI-1547229). The authors thank the participants of the Using Biodiversity Specimen-Based Data to Study Global Change workshop, hosted by iDigBio and the Missouri Botanical Garden, December 2015 (http://goo.gl/Q8APZH), for their contributions, and the three anonymous reviewers for their valuable comments.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

LITERATURE CITED

AIBS (American Institute of Biological Sciences). 2015a. Enhancing complex data integration across research domains: A workshop report. AIBS, Reston, Virginia, USA.

AIBS (American Institute of Biological Sciences). 2015b. Addressing biological informatics workforce needs: A report from the 2015 AIBS Council of Member Societies and Organizations Meeting. AIBS, Reston, Virginia, USA.

Anderson, R. P., M. Araújo, A. Guisan, J. M. Lobo, E. Martínez-Meyer, A. T. Peterson, and J. Soberón. 2016. Are species occurrence data in global online repositories fit for modeling species distributions? The case of the Global Biodiversity Information Facility (GBIF). Final report of the task group on GBIF data fitness for use in distribution modelling. GBIF, Copenhagen, Denmark.

Ariño, A. H., V. Chavan, and J. Otegui. 2016. Best practice guide for data gap analysis for biodiversity stakeholders. GBIF Secretariat, Copenhagen, Denmark.

Arnaud, E., N. P. Castañeda-Álvarez, J. G. Cossi, D. Endresen, E. Jahanshiri, and Y. Vigouroux. 2016. Final report of the Task Group on GBIF data fitness for use in agrobiodiversity. GBIF Secretariat, Copenhagen, Denmark.

August, T., M. Harvey, P. Lightfoot, D. Kilbey, T. Papadopoulos, and P. Jepson. 2015. Emerging technologies for biological recording. *Biological Journal of the Linnean Society* 115: 731–749.

Barnosky, A. D., E. A. Hadly, P. Gonzalez, J. Head, P. D. Polly, A. M. Lawing, and J. T. Eronen, et al. 2017. Merging paleobiology with conservation biology to guide the future of terrestrial ecosystems. *Science* 355(6325): eaah4787.

Bebber, D. P., M. A. Carine, J. R. I. Wood, A. H. Wortley, D. J. Harris, G. T. Prance, G. Davidse, et al. 2010. Herbaria are a major frontier for species discovery. *Proceedings of the National Academy of Sciences, USA* 107: 22169–22171.

Berendsohn, W. G., and P. Seltmann. 2010. Using geographical and taxonomic metadata to set priorities in specimen digitization. *Biodiversity Informatics* 7: 120–129.

Berents, P., M. Hamer, and V. Chavan. 2010. Towards demand-driven publishing: Approaches to the prioritization of digitization of natural history collection data. *Biodiversity Informatics* 7: 47–52.

Bishop, B. W., and C. Hank. 2016. Data curation profiling of biocollections. *Proceedings of the Association for Information Science and Technology* 53: 1–9.

Brummitt, N. A., S. P. Bachman, J. Griffiths-Lee, M. Lutz, J. F. Moat, A. Farjon, J. S. Donaldson, et al. 2015. Green plants in the red: A baseline global assessment for the IUCN sampled Red List Index for plants. *PLoS One* 10: e0135152.

Canhos, D. A. L., M. S. Sousa-Baena, S. de Souza, L. C. Maia, J. R. Stehmann, V. P. Canhos, R. De Giovanni, et al. 2015. The importance of biodiversity e-infrastructures for megadiverse countries. *PLoS Biology* 13: e1002204.

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

Casas-Marce, M., E. Revilla, M. Fernandes, A. Rodríguez, M. Delibes, and J. A. Godoy. 2012. The value of hidden scientific resources: Preserved animal specimens from private collections and small museums. *BioScience* 62: 1077–1082.

Cayuela, L., D. J. Golicher, A. Newton, M. Kolb, F. S. De Alburquerque, E. J. M. M. Arets, R. M. Alkemade, and A. M. Pérez. 2009. Species distribution modeling in the tropics: Problems, potentialities and the role of biological data for effective species conservation. *Tropical Conservation Science* 2: 319–352.

Chao, A., R. K. Colwell, C.-W. Lin, and N. J. Gotelli. 2009. Sufficient sampling for asymptotic minimum species richness estimators. *Ecology* 90: 1125–1133.

Chapman, A. D. 2005. Uses of primary species-occurrence data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen, Denmark.

Chapman, A. D., and J. Wieczorek [eds.]. 2006. Guide to best practices for georeferencing. Global Biodiversity Information Facility, Copenhagen, Denmark.

Chapman, A. D., and O. Grafton. 2008. Guide to best practices for generalising sensitive species-occurrence data, version 1.0. Global Biodiversity Information Facility, Copenhagen, Denmark.

Crawford, P. H. C., and B. W. Hoagland. 2009. Can herbarium records be used to map alien species invasion and native species expansion over the past 100 years? *Journal of Biogeography* 36: 651–661.

Creer, S., K. Deiner, S. Frey, D. Porazinska, P. Taberlet, W. K. Thomas, C. Potter, and H. M. Bik. 2016. The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution* 7: 1008–1018.

Culley, T. M. 2013. Why vouchers matter in botanical research. *Applications in Plant Sciences* 1(11): 1300076.

Daru, B. H., D. S. Park, R. B. Primack, C. G. Willis, D. S. Barrington, T. J. S. Whitfeld, T. G. Seidler, et al. 2018. Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytologist* 217: 939–955.

Davis, C. C., C. G. Willis, B. Connolly, C. Kelly, and A. M. Ellison. 2015. Herbarium records are reliable sources of phenological change driven by climate and provide novel insights into species' phenological cueing mechanisms. *American Journal of Botany* 102: 1599–1609.

Elith, J., and J. R. Leathwick. 2009. Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics* 40: 677–697.

Faith, D., B. Collen, A. H. Ariño, P. K. Patricia Koleff, J. Guinotte, J. Kerr, and V. Chavan. 2013. Bridging the biodiversity data gaps: Recommendations to meet users' data needs. *Biodiversity Informatics* 8: 41–53.

Gardner, J. L., T. Amano, W. J. Sutherland, L. Joseph, and A. Peters. 2014. Are natural history collections coming to an end as time-series? *Frontiers in Ecology and the Environment* 12: 436–438.

Geijzendorffer, I. R., E. C. Regan, H. M. Pereira, L. Brotons, N. Brummitt, Y. Gavish, and P. Haase, et al. 2016. Bridging the gap between biodiversity data and policy reporting needs: An Essential Biodiversity Variables perspective. *Journal of Applied Ecology* 53: 1341–1350.

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

Grinnell, J. 1912. An afternoon's field notes. *Condor* 14: 104–107.

Gueta, T., and Y. Carmel. 2016. Quantifying the value of user-level data cleaning for big data: A case study using mammal distribution models. *Ecological Informatics* 34: 139–145.

Guisan, A., R. Tingley, J. B. Baumgartner, I. Naujokaitis-Lewis, P. R. Sutcliffe, A. I. T. Tulloch, T. J. Regan, et al. 2013. Predicting species distributions for conservation decisions. *Ecology Letters* 16: 1424–1435.

Hampton, S. E., M. B. Jones, L. A. Wasser, M. P. Schildhauer, S. R. Supp, J. Brun, et al. 2017. Skills and knowledge for data-intensive environmental research. *BioScience* 67: 546–557.

Heberling, J. M., and B. L. Isaac. 2017. Herbarium specimens as exaptations: New uses for old collections. *American Journal of Botany* 104: 1–3.

Holmes, M. W., T. T. Hammond, G. O. U. Wogan, R. E. Walsh, K. LaBarbera, E. A. Wommack, F. M. Martins, et al. 2016. Natural history collections as windows on evolutionary processes. *Molecular Ecology* 25: 864–881.

Hortal, J., J. M. Lobo, and A. Jiménez-Valverde. 2007. Limitations of biodiversity databases: Case study on seed-plant diversity in Tenerife, Canary Islands. *Conservation Biology* 21: 853–863.

IPCC (Intergovernmental Panel on Climate Change). 2014. Climate Change 2014: Impacts, adaptation, and vulnerability. Part A: Global and sectoral aspects. Contribution of Working Group II to the fifth assessment report of the Intergovernmental Panel on Climate Change. Cambridge University Press, New York, New York, USA.

James, S. A. 2017. Publishing a new species? Add the unique identifiers! Research Spotlight: April 2017. Website https://www.idigbio.org/content/research-spotlight-april-2017 [accessed 9 September 2017].

Jarvis, A., A. Lane, and R. J. Hijmans. 2008. The effect of climate change on crop wild relatives. *Agriculture, Ecosystems & Environment* 126: 13–23.

Kohn, M. J., C. A. E. Strömberg, R. H. Madden, R. E. Dunn, S. Evans, A. Palacios, and A. A. Carlini. 2015. Quasi-static Eocene-Oligocene climate in Patagonia promotes slow faunal evolution and mid-Cenozoic global cooling. *Palaeogeography, Palaeoclimatology, Palaeoecology* 435: 24–37.

Krishtalka, L., E. Dalcin, S. Ellis, J. C. Ganglo, T. Hosoya, M. Nakae, I. Owens, et al. 2016. Accelerating the discovery of biocollections data. GBIF Secretariat, Copenhagen, Denmark.

Maguire, K. C., D. Nieto-Lugilde, J. L. Blois, M. C. Fitzpatrick, J. W. Williams, S. Ferrier, and D. J. Lorenz. 2016. A controlled comparison of species- and community-level models across novel climates and communities. *Proceedings of the Royal Society B* 283: 20152817.

McGeoch, M., Q. Groom, S. Pagad, V. G. Petrosyan, G. M. Ruiz, and J. Wilson. 2016. Task group on data fitness for use in research into invasive alien species. GBIF Secretariat, Copenhagen, Denmark.

Meyer, W. M., J. A. Eble, K. Franklin, R. B. McManus, S. L. Brantley, J. Henkel, P. E. Marek, et al. 2015. Ground-dwelling arthropod communities of a sky island mountain range in southeastern Arizona, USA: Obtaining a baseline for assessing the effects of climate change. *PLoS One* 10: e0135210.

Morales-Castilla, I., M. G. Matias, D. Gravel, and M. B. Araújo. 2015. Inferring biotic interactions from proxies. *Trends in Ecology and Evolution* 30: 347–356.

Morris, P., J. Hanken, D. Lowery, B. Ludäscher, J. Macklin, T. McPhillips, R. Morris, et al. 2017. Fitness-for-use-framework-aware data quality workflows in Kurator. *Proceedings of TDWG* 1: e20379.

Morrison, S. A., T. S. Sillett, W. C. Funk, C. K. Ghalambor, and T. C. Rick. 2017. Equipping the 22nd-century historical ecologist. *Trends in Ecology and Evolution* 32: 578–588.

Morueta-Holme, N., B. Blonder, B. Sandel, B. J. McGill, R. K. Peet, J. E. Ott, C. Violle, et al. 2016. A network approach for inferring species associations from co-occurrence data. *Ecography* 39: 1139–1150.

Murray, B. R., L. J. Martin, M. L. Phillips, and P. Pyšek. 2017. Taxonomic perils and pitfalls of dataset assembly in ecology: A case study of the naturalized Asteraceae in Australia. *NeoBiota* 34: 1–20.

Page, L. M., B. J. MacFadden, J. A. B. Fortes, P. S. Soltis, and G. Riccardi. 2015. Digitization of biodiversity collections reveals biggest data on biodiversity. *BioScience* 65: 841–842.

Phillips, R. D., A. P. Brown, K. W. Dixon, and S. D. Hopper. 2011. Orchid biogeography and factors associated with rarity in a biodiversity hotspot, the Southwest Australian Floristic Region. *Journal of Biogeography* 38: 487–501.

Pino-Del-Carpio, A., A. H. Ariño, A. Villarroya, J. Puig, and R. Miranda. 2014. The biodiversity data knowledge gap: Assessing information loss in the management of Biosphere Reserves. *Biological Conservation* 173: 74–79.

Poelen, J. H., J. D. Simons, and C. J. Mungall. 2014. Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecological Informatics* 24: 148–159.

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

Prendergast, J. R., S. N. Wood, J. H. Lawton, and B. C. Eversham. 1993. Correcting for variation in recording effort in analyses of diversity hotspots on JSTOR. *Biodiversity Letters* 1: 39–53.

Rauber, A., A. Asmi, van Uytvanck D., and S. Pröll. 2015. Data citation of evolving data. Website https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf [accessed 8 December 2017].

Robertson, M. P., V. Visser, and C. Hui. 2016. Biogeo: An R package for assessing and improving data quality of occurrence record datasets. *Ecography* 39: 394–401.

Rouhan, G., L. J. Dorr, L. Gautier, P. Clerc, S. Muller, and M. Gaudeul. 2017. The time has come for natural history collections to claim co-authorship of research articles. *Taxon* 66: 1014–1016.

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

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

Soberón, J., R. Jiménez, J. Golubov, and P. Koleff. 2007. Assessing completeness of biodiversity databases at different spatial scales. *Ecography* 30: 152–160.

Soltis, P. S. 2017. Digitization of herbaria enables novel research. *American Journal of Botany* 104: 1–4.

Soranno, P. A., E. G. Bissell, K. S. Cheruvelil, S. T. Christel, S. M. Collins, C. E. Fergus, C. T. Filstrup, et al. 2015. Building a multi-scaled geospatial temporal ecology database from disparate data sources: Fostering open science and data reuse. *GigaScience* 4: 28.

Strömberg, C. A. E., R. E. Dunn, R. H. Madden, M. J. Kohn, and A. A. Carlini. 2013. Decoupling the spread of grasslands from the evolution of grazer-type herbivores in South America. *Nature Communications* 4: 1478.

Suarez, A. V., and N. D. Tsutsui. 2004. The value of museum collections for research and society. *BioScience* 54: 69–74.

Sutherst, R. W. 2014. Pest species distribution modelling: Origins and lessons from history. *Biological Invasions* 16: 239–256.

Thiers, B. 2017. The World's herbaria 2016: A summary report based on data from Index Herbariorum. Website http://sweetgum.nybg.org/science/ih/ [accessed 9 September 2017].

Troudet, J., P. Grandcolas, A. Blin, R. Vignes-Lebbe, and F. Legendre. 2017. Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports* 7: 2045–2322.

Veiga, A. K., A. M. Saraiva, A. D. Chapman, P. J. Morris, C. Gendreau, D. Schigel, and T. J. Robertson. 2017. A conceptual framework for quality assessment and management of biodiversity data. *PLoS One* 12(6): e0178731.

Verheyen, K., P. De Frenne, L. Baeten, D. Waller, R. Hedl, M. Perring, J. Brunet, et al. 2017. Combining community resurvey data to advance global change research. *BioScience* 67: 73–83.

Walls, R. L., J. Deck, R. Guralnick, S. Baskauf, R. Beaman, S. Blum, S. Bowers, et al. 2014. Semantics in support of biodiversity knowledge discovery: An introduction to the biological collections ontology and related ontologies. *PLoS One* 9: e89606.

Ward, D. F., R. A. B. Leschen, and T. R. Buckley. 2015. More from ecologists to support natural history museums. *Trends in Ecology and Evolution* 30: 373–374.

Weirauch, C., K. C. Seltmann, R. T. Schuh, M. D. Schwartz, C. Johnson, M. A. Feist, and P. S. Soltis. 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.

Wieczorek, J., D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, T. Robertson, and D. Vieglais. 2012. Darwin Core: An evolving community-developed biodiversity data standard. *PLoS One* 7(1): e29715.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3: 160018.

Willis, C. G., E. R. Ellwood, R. B. Primack, C. C. Davis, K. D. Pearson, A. S. Gallinat, J. M. Yost, et al. 2017. Old plants, new tricks: Phenological research using herbarium specimens. *Trends in Ecology and Evolution* 32: 531–546.

Winston, J. E. 2007. Archives of a small planet: The significance of museum collections and museum-based research in invertebrate taxonomy. *Zootaxa* 1668: 47–54.

Zaharia, M., R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, et al. 2016. Apache Spark: A unified engine for big data processing. *Communications of the ACM* 59(11): 56–65.

