PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Review

Cite this article: Nelson G, Ellis S. 2018 The history and impact of digitization and digital data mobilization on biodiversity research. *Phil. Trans. R. Soc. B* **374**: 20170391. http://dx.doi.org/10.1098/rstb.2017.0391

Accepted: 8 August 2018

One contribution of 16 to a theme issue 'Biological collections for understanding biodiversity in the Anthropocene'.

Subject Areas:
bioinformatics, ecology, environmental science, evolution, plant science, taxonomy and systematics

Keywords:
digital data, biodiversity, data mobilization, digitization, Anthropocene, iDigBio

Author for correspondence:
Gil Nelson
e-mail: gnelson@bio.fsu.edu

The history and impact of digitization and digital data mobilization on biodiversity research

Gil Nelson[1] and Shari Ellis[2]

[1]iDigBio, Florida State University, 142 Collegiate Loop, Tallahassee, FL 32306, USA
[2]Florida Museum of Natural History, University of Florida, 1659 Museum Road, Gainesville, FL 32611, USA

GN, 0000-0002-7851-4445

The first two decades of the twenty-first century have seen a rapid rise in the mobilization of digital biodiversity data. This has thrust natural history museums into the forefront of biodiversity research, underscoring their central role in the modern scientific enterprise. The advent of mobilization initiatives such as the United States National Science Foundation's Advancing Digitization of Biodiversity Collections (ADBC), Australia's Atlas of Living Australia (ALA), Mexico's National Commission for the Knowledge and Use of Biodiversity (CONABIO), Brazil's Centro de Referência em Informação (CRIA) and China's National Specimen Information Infrastructure (NSII) has led to a rapid rise in data aggregators and an exponential increase in digital data for scientific research and arguably provide the best evidence of where species live. The international Global Biodiversity Information Facility (GBIF) now serves about 131 million museum specimen records, and Integrated Digitized Biocollections (iDigBio) in the USA has amassed more than 115 million. These resources expose collections to a wider audience of researchers, provide the best biodiversity data in the modern era outside of nature itself and ensure the primacy of specimen-based research. Here, we provide a brief history of worldwide data mobilization, their impact on biodiversity research, challenges for ensuring data quality, their contribution to scientific publications and evidence of the rising profiles of natural history collections.

This article is part of the theme issue 'Biological collections for understanding biodiversity in the Anthropocene'.

1. Introduction

Rapid technological advances during the Anthropocene have precipitated massive impacts on biodiversity as well as how biodiversity science is conducted. Because major shifts in the Earth's stratigraphy are primarily geological rather than cultural features, the Anthropocene Epoch can only be properly defined by how its stratigraphic signature differs from that of its immediate predecessor, the Late Holocene. Nevertheless, unlike earlier epochs, the Anthropocene is often characterized by non-geological but usually parallel human impacts, including the impacts on biodiversity that have resulted from massive increases in human population and the reciprocal impacts on humans, themselves. Potentially deleterious environmental impacts coupled with the rise and influence of digital technologies brought on by the Anthropocene have increased the urgency and tools for using museum specimens to enhance our understanding of biodiverse systems.

The previous two decades have seen exponential growth in the aggregation and availability of digital biodiversity data for use in research, conservation, outreach and integrated studies across all domains of the biodiversity sciences [1–8]. This has thrust natural history museums and academic collections—especially the biodiversity specimens they curate—into the forefront of biodiversity research in systematics [9], ecology and conservation, underscoring their central role in the modern scientific enterprise and making them more visible, accessible and

THE ROYAL SOCIETY PUBLISHING

transparent to citizen scientists and the general public. The advent of such digitization and data mobilization initiatives as the United States (US) National Science Foundation's Advancing the Digitization of Biodiversity Collections (ADBC) programme, Australia's Atlas of Living Australia (ALA), Mexico's Comisión Nacional Para el Conocimiento y Uso de la Biodiversidad (CONABIO), Brazil's Centro de Referência em Informação (CRIA), Europe's emerging Distributed System of Scientific Collections (DiSSCo) and China's National Specimen Information Infrastructure (NSII) has led to a rapid rise in regional, national and international digital data aggregators, and precipitated an exponential increase in the availability of digital data for scientific research. These digital resources raise the profiles of museums, expose collections to a wider audience of systematics and conservation researchers, provide the best biodiversity data outside of nature itself [10], ensure that natural history museums remain at the forefront of biodiversity science and open opportunities for addressing a litany of grand challenge questions [11].

Here, we provide a brief accounting of worldwide digital data generation and mobilization initiatives, the impact of these data on biodiversity research, challenges for improving and ensuring the quality of these data, new data underscoring the impact of worldwide digitization initiatives on scientific publications and evidence of the roles these activities play in raising the public and scientific profiles of natural history collections. Our primary focus is digitized museum specimens with only brief mention of ecological research data deposited in research repositories, expertly vetted range maps, satellite vegetation data or non-vouchered species observational data.

2. A history of digitization

Beginning with a 1999 recommendation of the Biodiversity Informatics Subgroup of the Organization for Economic Cooperation and Development's Megascience Forum, the Global Biodiversity Informatics Facility (GBIF) was founded to enable access to the vast quantities of biodiversity information to advance scientific research and increase knowledge of the natural world [12]. By mid-2018, GBIF was serving more than one billion biodiversity occurrence records, nearly 150 million (or 15%) of which were based on preserved specimens held in natural history collections. Concomitant with the establishment of GBIF and based on a recommendation of the Council of Heads of Australasian Herbaria (CHAH), the Australian Virtual Herbarium was created in 2001, the success of which led to funding for the ALA, a much broader initiative with the mission of transforming Australia's biodiversity knowledge into digital format for enabling collaboration in biodiversity research [13]. In the last decade, the ALA database has grown to over 73 million occurrence records, about 12.6 million (17.3%) of which represent preserved specimens.

Several countries in South America are also aggregating biodiversity data. Beginning in 2002, Brazil's CRIA launched the *species*Link [14] network with the goal of integrating species and specimen data available in natural history museums, herbaria and culture collections, and making these data openly and freely available on the Internet, along with tools to promote interoperability, integration, visualization and data cleaning [15,16]. As of January 2018, *species*Link served nearly 9 million records, about half of which are georeferenced, and at least some of which are also being served by GBIF as well as leading aggregators in North America. In 2010, the Brazilian government also launched ReFlora with the purpose of making available information on Brazilian plant specimens held in overseas herbaria. These data sources have become an important contributor to Brazilian conservation [17,18].

In November 2017, Mexico celebrated 25 years of its CONABIO, established in 1992 to promote, coordinate, support and carry out activities aimed at biodiversity knowledge, conservation and sustainability [19]. CONABIO is now serving nearly 6 million records through its World Biodiversity Information Network (REMIB), a large proportion of which are specimen records from natural history collections [20].

Asia, too, has moved forward with digitization and data mobilization activities. China's NSII is one of 28 initiatives funded by the country's Ministry of Science and Technology within the National Science and Technology infrastructure. NSII is designed to marshal data for use in conservation and the protection of China's biodiversity and serves as the GBIF node for China [21,22].

In Europe, the recently submitted proposal DiSSCo involves 21 European countries and 114 natural history museums with the stated mission of mobilizing, unifying and delivering 'bio- and geo-diversity information at the scale, form and precision required by scientific communities; transforming a fragmented landscape into a coherent and responsive research infrastructure' [23]. The project is centred at Naturalis Biodiversity Center, Leiden, The Netherlands and active work on the project is underway. If fully funded, the DiSSCo implementation timeline calls for deployment by 2024 [23].

Biodiversity specimen data digitization, mobilization and aggregation in the USA have been encouraged largely by the launch in 2011 of the US National Science Foundation's ADBC programme, its national resource, Integrated Digitized Biocollections (iDigBio) [24,25] and the several associated Thematic Collections Networks (TCN) [26], whose roles include generating and aggregating to iDigBio a wealth of digitized collections data to address grand challenge questions. To date, ADBC involves 708 collections in nearly 500 institutions representing all 50 of the US states and a majority of collection types [27]. Together, these institutions have contributed over 115 million text records and more than 26 million media records to the iDigBio portal [28]. Given that specimen object records often represent aggregated specimens stored in lots, trays, matrix or by collecting event, the number of individual physical specimens represented in these 115 million records is conservatively estimated at 300–400 million.

Worldwide and in parallel with or in some cases leading up to these national and international efforts, various larger museums with sufficient resources have been digitizing collections for at least two decades, serving data through institutional websites, with many now contributing data to leading aggregators. Examples from Europe include the Paris Herbarium, currently with about 5.4 million specimens digitized [29]; Natural History Museum, London, currently serving about 8.9 million specimen records [30–32]; Naturalis Biodiversity Center, The Netherlands, curating about 37 million objects, of which about 4 million have been digitized [33]; and Museum für Naturkunde in Berlin, with a major focus on whole-drawer digitization of insect trays [34]. The Global Plants Initiative [35,36], focused on making available type specimens of

plants, served as an important global leader in encouraging digitization. In the USA, the New York Botanical Garden (NYBG) [37], Harvard's Museum of Comparative Zoology (MCZ), the Harvard University Herbaria, the Yale Peabody Museum, Sam Noble Museum at the University of Oklahoma and the Museum of Vertebrate Zoology (MVZ) at University of California, Berkeley, the latter of which computerized its specimen data in the late 1970s to early 1980s and made them available online in 1997, were among the earliest digitizing institutions. MVZ was also a leader in the establishment of VertNet [38,39], a combination of several discipline-specific sub-projects and an early leader in the development of workflows, data quality protocols and label digitization standards. FishNet, now FishNet 2, was an early collaborator with the consortium that launched VertNet and has been a leader in the development of standards and protocols for georeferencing as well as an important aggregator for fish specimen data.

Despite the worldwide increase in digitization activities, there remain important regions that are poorly represented. Perhaps chief among these is Russia, which has large quantities of biodiversity data stored mostly in local databases inaccessible to the Internet [40]. Nevertheless, a number of Russia-based digitization projects have been launched, with the expectation that more will follow [40]. The continent of Africa is also moving forward with digitization under the auspices and encouragement of the South African National Biodiversity Institute (SANBI) and GBIF. Beginning with the development of a mobilization strategy in 2013–2015, SANBI has recently launched The African Biodiversity Challenge to facilitate data mobilization in Rwanda, Ghana, Malawi and Namibia [41]. Biodiversity information for India is being tracked in several national databases [42], but there are still large gaps in availability, especially of specimen-based digital data generated from Indian collections. As of 13 June 2018, the iDigBio portal contained approximately 361 000 records of Indian specimens, nearly all of which are from US and UK institutions. In 2008, India established the India Biodiversity Data Portal, which serves a variety of species, maps and related data, including over 1 million observation records. Given India's history as an important collecting destination for at least three centuries, there is growing interest in digitizing Indian specimens that are held in museums outside as well as inside India [43].

3. Digitization definition and approaches

We define digitization as the conversion of specimen data from analogue to digital signals. This includes transcribing text data from specimen labels and other specimen-related documents into digital records of those labels and documents regardless of input mode (e.g. voice, keyboard, scanning/optical character recognition (OCR)); the translation of physical specimens to digital images of those specimens, including two-dimensional, three-dimensional (3D), computed tomography (CT) and other digital image types that visually represent the physical specimen; the conversion of analogue audio and video recordings to digital recordings; the conversion of textual location descriptions into digital georeferences within an accepted geographical coordinate system and the conversion of other specimen-related data into digital format with technologies that are or might become available. Although in common parlance, some observers use 'digitize' to mean imaging and 'databasing' to mean text transcription, here we use digitization to encompass both.

Approaches to digitization and the workflows that flow from them vary by institution based on institutional goals, resources, personnel, curator preferences and collection types [44]. Nelson *et al.* [2] outline five digitization task clusters in common use. These clusters have provided guidance for the development of several workflow documents [45–47] that encompass numerous discipline-specific approaches to digitization protocols.

Embedded within virtually all approaches to digitization is the adherence to data standards that govern the elements to be included in text transcription and multimedia resources metadata. Essentially all biodiversity databases and major aggregators are designed around or provide methods for translation to Darwin Core [48], the most common and complete vocabulary for biodiversity data. Likewise, Audubon Core [49] provides standards for multimedia resources associated with specimens. These standards provide translation to a common language, making possible comparisons across data stores and disciplines.

4. Growth of digitization

The rapid increase in the generation and mobilization of digital data and the attractiveness of these data to biodiversity scientists have been paralleled by an equally rapid and upward trending number of publications using and referencing the output of numerous digitization projects. For example, since the inception of ADBC in 2011, there has been a steady rise in the number of publications that cite use of data and other resources (e.g. geographical coordinates) from the iDigBio aggregation portal, TCN portals or other portals that aggregate TCN data (figure 1). Moreover, while the number of publications authored by those funded by ADBC has been relatively constant, the number of publications by authors external to the ADBC community has shown a dramatic increase (figure 2). We take these increasing numbers as evidence of the value that biodiversity scientists and researchers attribute to the growing accumulation of digital data.

5. Research with digitized specimen data

Expanding availability of digital data is enhancing avenues for current and future research that stretches across the various domains represented in the neo- and palaeobiological sciences. For example, Soltis & Soltis [4] outline several emerging big data tools for analysing the increasingly large biodiversity datasets that are rapidly coming online, and suggest novel research questions these data might address. Research emphases include assessing phylogenetic diversity for conservation [50], large platform tools for integrated geospatial analyses using specimen locality data, advances in ecological niche and species distribution modelling [51–53], and the potential development of new workflows [4]. Losos *et al.* [11] have suggested how the burgeoning supply of digitized data might be used to address important human issues, including evolutionary medicine, food security, biodiversity sustainability, computation and design, evolution and justice, and the development of new types of biodiversity theories that accommodate newly emerging data streams.

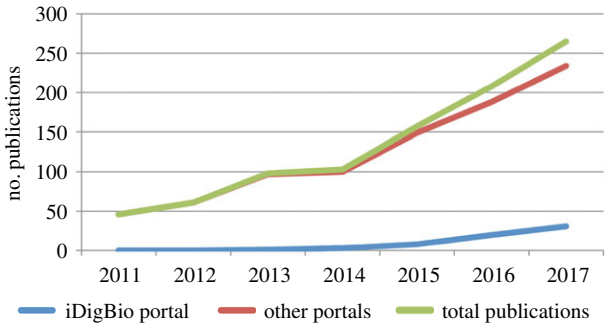

Figure 1. Publications that acknowledged searching, accessing or downloading data from one or more of the following portals: idigbio.org, lichenportal.org, bryophyteportal.org, tcn.amnh.org, symbiota4.acis.ufl.edu/scan/portal/, neherbaria.org, mycoportal.org, macroalgae.org, macaulaylibrary.org, seinet.org, sernecportal.org, vertnet.org, midatlanticherbaria.org, invertebase.org. The data supporting this figure and figure 2 can be accessed at https://www.idigbio.org/sites/default/files/internal-docs/Supporting%20references%20for%20Nelson%20%26%20Ellis%20%282018%29.pdf.

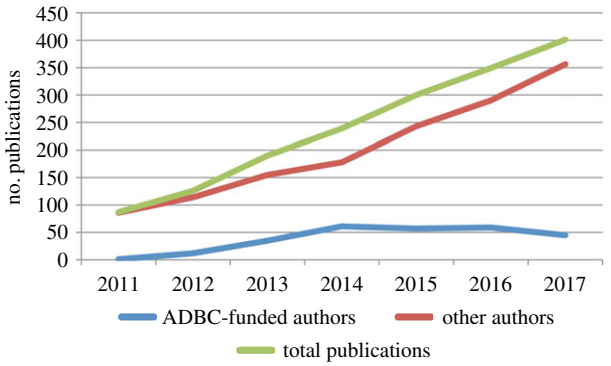

Figure 2. Publications (i) that include at least one author receiving ADBC funding as indicated by inclusion of grant number, affiliation as listed on byline and/or on NSF Award page; (ii) that mention ADBC supported projects including DigBio or any TCN; and (iii) use data from idigbio.org or any of the portals listed in figure 1.

Others have addressed emerging research angles, including the supplementation of existing datasets with related digital layers to enhance niche and species distribution modelling [54]; the use of 3D data for generating and testing new hypotheses; the implementation of convolutional neural networks (CNN) and deep learning in the analysis of image data for taxonomic determination [55–57] and specimen curation [58], the delineation of traits in specimen images and the determination and identification to genus or species of sediment-deposited pollen grains [59].

The delineation of traits in specimen images can be especially useful for detecting and relating phenological shifts in the fruiting and flowering times of vascular plants to the dynamics of climate change and the synchronicity of fruit production to wildlife migration (see Deacy *et al.* [60] for an example of where this could be applied). Phenology has also become an important exemplar for the study and tracking of global change [61–66], especially in the use of digital herbarium records for the study and tracking of phenological shifts in vascular plants [65,66] and fungi [67,68]. New tools and protocols are being advanced for rapid digitization [69] and automated scoring of herbarium sheets [70] and improved crowdsourcing platforms developed [71–74] that can be used to engage public participation in scoring specimen images for phenological stage.

Building on the rich history of using plant specimens to study phenology [66], the Phenology Working Group [75], hosted by iDigBio, has so far conducted one workshop resulting in two papers [66,70], is producing a special issue of *Applications in Plant Sciences* devoted to phenology and herbarium data, hosted a symposium on phenology and digital data at Botany 2018 and is currently researching the use of CNN in deep learning for mass scoring of specimen images using computer vision techniques. Part of the working group's interest lies in the synchronization of plant phenological stages with food availability to wildlife, an issue that has been demonstrated to influence wildlife behaviour and adaptation [60,76]. The potential for CNN impact on agriculture and food security for humans is also being demonstrated [77] and presents another avenue for promising research in the face of global change.

Within the last 3–5 years, the use of CT scanning has advanced from being applied mostly to fossils to a much wider range of specimens. The recently US National Science Foundation (NSF)-funded openVertebrate (oVert) TCN [78] is an example. oVert is using CT technology to scan 20 000 fluid-preserved vertebrate specimens, representing approximately 80% of living vertebrate genera. These specimens include fluid-preserved birds, reptiles, amphibians, caecilians, fishes and mammals. This collaboration of 18 institutions is the first TCN to provide the international research community with freely accessible digital 3D data for internal anatomy across vertebrate diversity. When applied to research, these types of data facilitate the study of patterns of relationships among living and extinct vertebrates, allow testing of hypotheses related to morphological evolution and adaptation, and promote the exploration of relationships between brain and nervous system anatomy as well as sensory and musculoskeletal function, all of which have the potential for significantly improving the human condition.

CT technology has also been used with Echinoides to explain the strength of lightweight skeletal structures [79]. Discoveries from these studies have provided 'the potential to improve technical multi-plated, lightweight and load-bearing structures for civil engineering, which make them valuable role models for structural analyses' [79, p. 6]. Such extrapolations from the study of biodiversity to other domains suggest implications of specimen-based research for the development of low-cost housing, food security [77] and medicine.

Other emerging opportunities include the layering of various environmental, ecological, behavioural, audio, visual and well-vetted observational datasets (such as those of the Cornell Laboratory of Ornithology's eBird project [80]) with digital specimen data to facilitate triangulation of multiple data sources as well as richer research methodologies and outcomes. Recent research [81], for example, has combined historical precipitation data with digitized museum records to correlate the well-documented periodic emergences of cicada populations with rain patterns to predict future emergences. Emergence events are clearly documented in specimen collection records, making them an excellent subject for combining these types of datasets.

Vertebrate zoologists are also finding ways to leverage multiple digital datasets. In 2013, NSF funded the Developing a Centralized Archive of Vouchered Animal Communication Signals TCN [82]. This collaboration of seven institutions led by researchers at The Cornell Laboratory of Ornithology is a

first step in expanding the scope of specimen-based research as well as broadening the definition of *specimen*. Macaulay Library [83], an international resource of nearly 6 million photo, audio and video objects (which also include arthropods), promotes the linking of physical vouchers with media records to provide foundational and coordinated datasets for studying the tempo and mode of animal signal evolution. As part of their research, the team is exploring a re-definition of *specimen* to include an extended suite of data, sometimes in the absence of physical objects [84]. Geospatial, temporal and phylogenetic analyses of digital specimen data have also been used for testing and reconciling controversial tenets and predictions of mimicry theory between coral snakes and other red-black banded snakes [85].

A recently established working group, also led by researchers at Cornell, is exploring methods for efficiently scoring, standardizing, analysing and presenting behavioural and movement data, such as those generated from camera traps, audio recording devices and extensive video studies of phenotypic behaviour. To date, two workshops have been held that combined behavioural scientists with data storage, analysis and aggregation experts (G. Nelson 2016, personal observation). Publications from these workshops are in progress. Behavioural data, such as those used by Brainerd, are resulting in increased understanding of the relationships between anatomy, morphology and biomechanics, including novel applications for assessing the biomechanics of birds in flight [86]. For analysing these types of data, Brainerd *et al.* [87] at Brown University have developed X-ray Reconstruction of Moving Morphology (XROMM), a 3D imaging technology for visualizing rapid skeletal movement *in vivo*.

For some museum scientists, the use of non-verifiable observational data is anathema to research that is dependent on physical vouchers, reliable and reproducible species identifications, and procedural replication. However, when well-vetted observations are used to supplement specimen data or foster the collection of new physical or media vouchers to test hypotheses, arguments against augmenting or enriching physical specimen datasets with observational datasets become less compelling. This is especially true in vertebrate zoology, where image or audio data are nearly as good as a specimen in hand for some types of research and is one of the underpinning themes of Webster and colleagues [84,88]. In addition, Peterson and his team [89–91] have combined carefully cleaned specimen and observational data from GBIF, VertNet, REMIB, Unidad de Informática para la Biodiversidad and eBird as well as other vetted sources to study extinctions, range shifts, phenological shifts and breakdown of interactions in ecological communities in the USA and Mexico over several decades.

6. Caveats

Digital data proliferation has revealed challenges as well as opportunities, especially with ensuring that aggregated data reflect the basic definition of quality, meaning that the data are complete, consistent, accurate, fit for use, free of bias [92–95] and adhere to community-embraced standards (e.g. the Darwin Core Standard [48]) [96–98]. The critical need for enhancing data quality has led to procedures, research methods and best practices for improving and confirming accuracy and fitness [97,99], including the combining of GBIF and GenBank data to identify potential identification anomalies in mycology [100], address pressing data quality challenges in entomology [96,101], mining and analysing palaeobiology data [102], discovering research uses for vertebrate trait data [103], reviewing and critiquing the efficacy and potential bias in species distribution models using natural history museum specimen data [52], combining El Niño–Southern Oscillation and 100 years of museum specimen data for the prediction of cicada emergence in Western North America [81] and the use of images to detect new ant host species for a common parasite [104]. Issues with data completeness have been documented in several studies (e.g. [105]), especially where gaps in distribution do not reflect expectations, suggesting under collecting or an equally likely dearth of mobilized records from one or more significant biodiversity collections.

Two major areas of improvement in the quality of digital data include the resolution and correction of taxon names as reported in electronic records of specimen label data [98,99] and the accuracy, resolution and fitness for use of reported geospatial coordinates [97,98]. Chapman [99] highlights three main types of taxon name errors, those of identification, spelling and format. Zermoglio *et al.* [106] add to this list errors that arise from misunderstanding, misapplication or lack of following the Darwin Core Standard, and highlight the use of out-of-date synonyms as problematic. Several projects have tackled the taxonomy and synonymy issue, but comprehensive solutions are few, with the possible exceptions of ornithology where worldwide recommendations of common names have a long history, and ichthyology, where the Catalog of Fishes [107] serves as the standard for nomenclature and taxonomy. In the long run, successful integration across the universe of digitized specimens, with the ultimate goal of linking specimen records to all of their derivatives (e.g. tissues, traits, genetic sequences and field notes) and commonalities across the Internet, including locality and taxonomic descriptions, temporally and spatially related specimens, directly and indirectly related literature, associated media records (e.g. audio and video recordings as well as still images of a specimen and its collecting site) and a potential host of other related information, is likely to be as dependent on well-ordered and fully documented digital systems for resolving taxonomy and nomenclature as it will be on the effective assignment of globally unique identifiers and semantic tags to specimen records. However, whether there will ever be widely accepted and incontrovertible taxonomies is somewhat conjectural. Taxonomy as a hierarchy of hypotheses is central to biodiversity science and to the scientific enterprise. Varying interpretations are to be expected.

For typical errors with geospatial data, Hill *et al.* [97] emphasize incomplete coordinates, strings inserted into numeric fields, incorrect coordinate system references, latitude values incorrectly reported for longitude and vice versa, incorrect or omitted numerical signs, misplaced decimals and coordinate values beyond a valid range. Aggregators have implemented tools to filter and correct, or at least suggest corrections for, some of these errors. However, errors in precision based on the quality of the global positioning system device used, georeferencing protocols, transcription errors, rounding and conversions from United States Public Land Survey System references to geographical coordinates can be much more troublesome, especially in studies where highly resolved coordinates are required. Append to these the assignment of coordinates to legacy records post collection, where georeferencers often make assignments from sparse descriptions on labels, and the opportunity for error is apparent.

7. Raising the public profiles of natural history museums and academic collections

Evidence suggests that the broad access to digital data over the last decade has contributed significantly to the public profiles of natural history museums and academic collections, at least in the USA. Reflecting a worldwide trend [7], one of the US National Science Foundation's underpinning goals for establishing the ADBC programme has been to raise the visibility of natural history museums by making them more accessible to school-age children, natural history enthusiasts and the public at large, educating these audiences that museums are not only interpretive organizations with exhibits and displays, but also significant research institutions that foster important discoveries and advances in our understanding and conservation of biodiversity [107]. Using current technologies to make natural history collections remotely accessible to a far wider audience has served to enhance research diversity [108] and elevate collections in ways that have fostered their increasing presence in the popular press made their contributions explicit and transparent. Although we admit that the conclusions we draw regarding raising museum profiles are anecdotal and not founded on extensive surveys or comprehensive and comparative scoring of popular press articles over time, we believe such quantitative investigations to be worthy of future research efforts.

A 2015 story in the *New York Times* [34] underscored the importance of getting museum data online and a companion article offered a guide to five digital resources that offer access to natural history collections [109]. In February 2017, the *Washington Post* published a video [110] entitled 'These three people, and one conveyor belt, are digitizing millions of plant specimens', highlighting the work being done at the herbarium of the Smithsonian's National Museum of Natural History (NMNH) in Washington, D.C., which houses a collection of about 5 million dried and mounted plant specimens. In September 2017, the *Chicago Tribune* highlighted the importance of collections in a video entitled 'Endangered Insects at the Field Museum [111]'. NMNH and the Field are two of the United States' largest and best-known museums.

In 2016, *Voice of America News* featured digitization efforts underway at the Natural History Museum of Los Angeles County [112], reinforcing the notion that digitizing the huge numbers of specimens in natural history collections will facilitate discovery by making specimen searches and comparisons more efficient and timely. The Canadian Museum of Nature was highlighted in a 2014 *CBCNews* feature [113] for their programme to digitize 3 million of their 10 million specimens, in what was presumably the museum's first round of digitization activities.

In some instances, the popular press includes citations or links to the original scientific papers that the popular article intends to interpret, such as the paper by ter Steege *et al.* [114] which was reported on in the *Science* section of the *New York Times* on 13 July 2016 [115]. The article highlighted the use of digital records to construct an inventory of Amazonian trees [116]. In times when budgets and support for collections seem to be declining, provocative titles like 'What can you do with 300,000 dead bees?', which appeared in the *Toronto Star*, 25 January 2016 [117] heading an article regarding the importance of the bee collection at the Royal Ontario Museum, make visible and lend transparency to the important science achieved through the maintenance of natural history museums and their specimens.

The elevated profile of natural history museums as holders of biodiversity specimens and the digital data that represent them, in addition to interpretive kiosks and displays, has not been lost on undergraduate students, who themselves become outreach agents [118]. As museums reach out even more aggressively, exposing undergraduates to collections-based research and the incorporation of digital data in biodiversity science, the potential for downstream impacts, including recruitment of a more diverse constituency and a broader range of skill sets, will grow [119,120].

8. Conclusion

The increasing pace of digital specimen data mobilization coupled with the rapid development of tools and protocols for the novel use of these data have placed natural history museums and herbaria at the forefront of biodiversity research, increasing their visibility and undergirding their value to scientists and the general public. Enhanced opportunities for research and data analysis are leading to discoveries across all biodiversity domains as well as informing research in engineering, design, architecture, food security and the medical sciences. The recent expansion of digital data has placed biodiversity collections on the cusp of big data science, opening multiple pathways for natural history museums to make positive contributions to our understanding of and responses to impending global change.

Data accessibility. The data supporting figures 1 and 2 can be accessed at: https://www.idigbio.org/sites/default/files/internal-docs/Supporting%20references%20for%20Nelson%20%26%20Ellis%20%282018%29%20updated.pdf.

Authors' contributions. G.N. made substantial contributions to the design of the manuscript, drafted the original manuscript, revised it as needed and approved the final version. S.E. compiled, formatted and analysed data, critically reviewed and improved the manuscript, and approved the final version. G.N. has collaborated with the guest editor Meineke on symposia and has invited her to contribute a paper for a special edition of *Applications in Plant Sciences*.

Competing interests. We declare we have no competing interests.

Funding. The authors are wholly or partially funded by US National Science Foundation award DBI 1547229.

References

1. Smith VS, Blagoderov V. 2012 Bringing collections out of the dark. *ZooKeys* **209**, 1–6. (doi:10.3897/zookeys.209.3699)
2. Nelson G, Paul D, Riccardi G, Mast A. 2012 Five task clusters that enable efficient and effective digitization of biological collections. *ZooKeys* **209**, 19–45. (doi:10.3897/zookeys.209.3135)
3. Page LM, MacFadden BJ, Fortes JA, Soltis PS, Riccardi G. 2015 Digitization of biodiversity collections reveals biggest data on biodiversity. *BioScience* **65**, 841–842. (doi:10.1093/biosci/biv104)
4. Soltis DE, Soltis PS. 2016 Mobilizing and integrating big data in studies of spatial and phylogenetic patterns of biodiversity. *Plant Diversity* **38**, 264–270. (doi:10.1016/j.pld.2016.12.001)
5. Cantrill D 2018 The Australasian Virtual Herbaria: tracking data usage and benefits for biological

collections. *Appl. Plant Sci.* **6**, e1026. (doi:10.1002/aps3.1026)

6. Howarth F 2018 The future of research in natural history museums. In *The future of natural history museums* (ed. E Dorfman), pp. 65–81. London, UK: Routledge, Taylor and Francis Group.
7. Norris C. 2018 The future of natural history collections. In *The future of natural history museums* (ed. E Dorfman), pp. 13–28. London, UK: Routledge, Taylor and Francis Group.
8. Soltis PS, Nelson G, James SA. 2018 Green digitization: online botanical collections data answering real-world questions. *APPS* **6**. (doi:10.1002/aps3.1028)
9. Ang Y *et al.* 2013A plea for digital reference collections and other science-based digitization initiatives in taxonomy: sepsidnet as exemplar. *Syst. Entomol.* **38**, 637–644. (doi:10.1111/syen.12015)
10. Page L. 2014 Digitization in the Pacific. Presented at Biological Digitization in the Pacific Workshop, Honolulu, pp. 25–27, March 2014. See https://www.idigbio.org/wiki/index.php/PacificDigitization (accessed 21 February 2018).
11. Losos JB *et al.* 2013 Evolutionary biology for the 21st century. *PLoS Biol.* **11**, e1001466. (doi:10.1371/journal.pbio.1001466)
12. GBIF. 2018 What is GBIF? See https://www.gbif.org/what-is-gbif (accessed 5 January 2018).
13. ALA. 2018 Who we are. See https://www.ala.org.au/who-we-are/ (accessed 5 January 2018).
14. speciesLink. 2018 *species*Link *The* Project. See http://splink.cria.org.br/description?criaLANG=en (accessed 5 January 2018).
15. CRIA. 2018 *species*Link. See http://splink.cria.org.br/index?&setlang=en (accessed 17 January 2018).
16. Sousa-Baena MS, Garcia LC, Peterson AT. 2013 Knowledge behind conservation status decisions: data basis for 'Data Deficient' Brazilian plant species. *Biol. Conserv.* **173**, 80–89. (doi:10.1016/j.biocon.2013.06.034)
17. Forzza RC *et al.* 2012 New Brazilian floristic list highlights conservation challenges. *BioScience* **62**, 39–45. (doi:10.1525/bio.2012.62.1.8)
18. Davis CC, Ellison AM. 2018 The brave new world of the digital herbarium mobilizing the past to understand the future. *ReVista*, Spring 2018. See https://revista.drclas.harvard.edu/book/brave-new-world-digital (accessed 10 August 2018)
19. Comisión Nacional para el Conocimiento y Uso de la Biodiversidad. 2017 CONABIO, 25 años de evolución. See https://www.gob.mx/cms/uploads/attachment/file/262393/25_an_os_Conabio_web.pdf (accessed 17 January 2018)
20. REMIB. 2018 See http://www.conabio.gob.mx/remib/cgi-bin/clave_remib.cgi?lengua=EN (accessed 17 January 2018)
21. NSII. 2018 About us. See http://www.nsii.org.cn/2017/AboutUs-en.php (accessed 18 January 2018)
22. NSII. 2018 See http://www.nsii.org.cn/2017/New.php?node=125 (accessed 18 January 2018)
23. DiSSCo. 2018 What is DiSSCo? See http://dissco.eu/ (accessed 18 January 2018)
24. AIBS. 2012 Implementation plan for the network integrated biocollections alliance. See https://www.nsf.gov/bio/pubs/reports/niba_implementation_plan.pdf (accessed 25 February 2018)
25. NSF. 2018 Advancing digitization of biodiversity collections (ADBC). See https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503559 (accessed 17 January 2018)
26. iDigBio. 2018 TCNs. See https://www.idigbio.org/wiki/index.php/TCNs (accessed 17 January 2018)
27. iDigBio. 2018 Collaborating Institutions. See https://www.idigbio.org/content/collaborating-institutions (accessed 17 January 2018)
28. iDigBio. 2018 Welcome to the iDigBio Portal. See https://www.idigbio.org/portal (accessed 12 February 2018)
29. Le Bras G *et al.* 2017 The French Muséum national d'histoire naturelle vascular plant herbarium collection dataset. *Sci. Data* **4**. (doi:10.1038/sdata.2017.16)
30. NHM. 2018 Digital collections programme. See http://www.nhm.ac.uk/our-science/our-work/digital-museum/digital-collections-programme.html (accessed 18 Jan 2018)
31. NHM. 2018 Data portal. See http://data.nhm.ac.uk/?_ga=2.137559288.999941569.1516284951–220695054.1516284951 (accessed 18 January 2018)
32. Blagoderov V, Kitching IJ, Livermore L, Simonsen TJ, Smith VS. 2012 No specimen left behind: industrial scale digitization of natural history collections. *ZooKeys* **209**, 133–146. (doi:10.3897/zookeys.209.3178)
33. Naturalis. 2018 Digitization. See https://science.naturalis.nl/en/collection/digitization/ (accessed 18 January 2018)
34. Olsen E. 2015 Museum specimens find new life online. *New York Times, Science*. See https://www.nytimes.com/2015/10/20/science/putting-museums-samples-of-life-on-the-internet.html (accessed 18 January 2018)
35. Smithsonian Tropical Research Institute. 2011 Online access to the plants of the world is available. EurekAlert! See https://www.eurekalert.org/pub_releases/2011-01/stri-oat011111.php (accessed 18 January 2018)
36. Funk VA. 2013 The Global Plants Initiative celebrates its achievements and plans for the future. *Taxon* **62**, 417–418.
37. Thiers BM, Tulig MC, Watson KA. 2016 Digitization of the New York Botanical Garden Herbarium. *Brittonia* **68**, 324–333. (doi:10.1007/s12228-016-9423-7)
38. Constable H *et al.* 2010 VertNet: a new model for biodiversity data sharing. *PLoS Biol.* **8**, e1000309. (doi:10.1371/journal.pbio.1000309)
39. Guralnick R, Constable H. 2010 VertNet: creating a data-sharing community. *BioScience* **60**, 258–259. (doi:10.1525/bio.2010.60.4.2)
40. Ivanova NV, Shashkov MP. 2016 Biodiversity databases in Russia: towards a national portal. *Arctic Science*. See http://www.nrcresearchpress.com/doi/pdfplus/10.1139/AS-2016-0050 (accessed 9 August 2018)
41. SANBI. 2018 The African Biodiversity Challenge. See http://biodiversityadvisor.sanbi.org/participation/mobilising-africas-biodiversity-data/the-african-biodiversity-challenge/ (accessed 18 January 2018)
42. Vattakaven T *et al.* 2016 India biodiversity portal: an integrated, interactive and participatory biodiversity informatics platform. *Biodivers. Data J.* **4**, e10279 (doi:10.3897/BDJ.4.e10279)
43. Pavid K. 2017 *Thousands of Indian plants to be digitised for the first time*. Natural History Museum. See http://www.nhm.ac.uk/discover/news/2017/november/thousands-of-indian-plants-to-be-digitised-for-the-first-time.html (accessed 9 August 2018)
44. Smith V, Blagoderov V (eds). 2012 No specimen left behind: mass digitization of natural history collections. Special Edition, ZooKeys **209**.
45. Nelson G *et al.* 2015 Digitization workflows for flat sheets and packets of plants, algae, and fungi. *Appl. Plant Sci.* **3**, 1500065. (doi:10.3732/apps.1500065)
46. Karim TS, Burkhalter R, Farrell UC, Molineux A, Nelson G, Utrup J, Butts S. 2016 Digitization workflows for paleontology collections. *Palaeontol. Electron.* **19.3.4T**, 1–14.
47. iDigBio. 2018 Workflow modules and task lists. See https://www.idigbio.org/content/workflow-modules-and-task-lists) (accessed 16 June 2018)
48. Wieczorek J, Döring M, De Giovanni R, Robertson T, Vieglais D. 2009 Darwin Core Terms: a quick reference guide. See http://rs.tdwg.org/dwc/terms (accessed 23January 2018)
49. Biodiversity Standards. 2018 Audubon core. See https://terms.tdwg.org/wiki/Audubon_Core (accessed 19 June 2018)
50. Mishler BD, Knerr N, González-Orozco CE, Thornhill AH, Laffan SW, Miller JT. 2014 Phylogenetic measures of biodiversity and neo- and paleo-endemism in Australian Acacia. *Nat. Commun.* **5**, 4473. (doi:10.1038/ncomms5473)
51. Phillips SJ, Anderson RP, Schapire RE. 2005 Maximum entropy modeling of species geographic distributions. Ecological modelling. *Ecol. Model.* **190**, 231–259. (doi:10.1016/j.ecolmodel.2005.03.026)
52. Elith J, Leathwick JR. 2009 Species distribution models: ecological explanation and prediction across space and time. *Annu. Rev. Ecol. Evol. Syst.* **40**, 677–697. (doi:10.1146/annurev.ecolsys.110308.120159)
53. Gomes VH *et al.* 2018 Species distribution modelling: contrasting presence-only models with plot abundance data. *Sci. Rep.* **8**, 1–12. (doi:10.1038/s41598-017-18927-1)
54. Visger CJ, Germain-Aubrey CC, Patel M, Sessa EB, Soltis PS, Soltis DE. 2016 Niche divergence between diploid and autotetraploid *Tolmiea*. *Am. J. Bot.* **103**, 1396–1406. (doi:10.3732/ajb.1600130)
55. Carranza-Rojas J, Goeau H, Bonnet P, Mata-Montero E, Joly A. 2017 Going deeper in the automated identification of herbarium specimens. *BMC Evol. Biol.* **17**, 181. (doi:10.1186/s12862-017-1014-z)
56. Botella C, Joly A, Bonnet P, Monestiez P, Munoz F. 2018 Species distribution modeling based on the

automated identification of citizen observations. *APPS* **6**. (doi:10.1002/aps3.1029)

57. Younis S, Weiland C, Hoehndorf R, Dressler S, Hickler T, Seegar B, Schmidt M. 2018 Taxon and trait recognition from digitized herbarium specimens using deep convolutional neural networks. *Botany Lett.* 1–7.
58. 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. *Biodivers. Data J.* **5**, e21139. (doi:10.3897/BDJ.5.e21139)
59. Haselhorst D. 2017 Using convolutional neural networks to automate tropical pollen counts and identification. *Research Spotlight*. September 2017. See https://www.idigbio.org/content/research-spotlight-september-2017 (accessed 13 March 2018)
60. Deacy WW, Armstrong JB, Leacock WB, Robbins CT, Gustine DD, Ward EJ, Erlenbach JA, Stanford JA. 2017 Phenological synchronization disrupts trophic interactions between Kodiak brown bears and salmon. *Proc. Natl Acad. Sci. USA* **114**, 10432–10 437. (doi:10.1073/pnas.1705248114)
61. Gallinat AS, Russo L, Melaas E, Willis CG, Primack R. 2018 Herbarium specimens show patterns of fruiting phenology in native and invasive plant species across New England. *Am. J. Bot*. **105**, 31–41. (doi:10.1002/ajb2.1005)
62. Meineke EK, Davies TJ, Daru BH, Davis CC. 2018 Biological collections for understanding biodiversity in the Anthropocene. *Phil. Trans. R. Soc. B* **374**, 20170386. (doi:10.1098/rstb.2017.0386)
63. Park DS, Breckheimer I, Williams AC, Law E, Ellison AM, Davis CC. 2018 Herbarium specimens reveal substantial and unexpected variation in phenological sensitivity across the eastern United States. *Phil. Trans. R. Soc. B* **374**, 20170394. (doi:10.1098/rstb.2017.0394)
64. Davis CC, Willis CG, Connolly B, Courtland K, Ellison AM. 2015 Herbarium records are reliable sources of phenological change driven by climate and provide novel insights into species' phenological cueing mechanisms. *Am. J. Botany* **102**, 1599–1609. (doi:10.3732/ajb.1500237)
65. Willis CG *et al.* 2017 CrowdCurio: an online crowdsourcing platform to facilitate climate change studies using herbarium specimens. *New Phytol.* **215**, 479–488. (doi:10.1111/nph.14535)
66. Willis CG *et al.* 2017 Old plants, new tricks: phenological research using herbarium specimens, *Trends Ecol. Evol.* **32**, 531–546. (doi:10.1016/j.tree.2017.03.015)
67. Andrew C, Heegaard E, Gange AC, Senn-Irlet B, Egli S, Kirk PM, Büntgen U, Kauserud H, Boddy L. 2018 Congruency in fungal phenology patterns across dataset sources and scales. *Fungal Ecol.* **32**, 9–17. (doi:10.1016/j.funeco.2017.11.009)
68. Andrew C *et al.* 2017 Big data integration: Pan-European fungal species observations' assembly for addressing contemporary questions in ecology and global change biology. *Fungal Biol. Rev.* **31**, 88–98. (doi:10.1016/j.fbr.2017.01.001)
69. Sweeney PW, Starly B, Morris PJ, Xu Y, Jones A, Radhakrishnan S, Grassa CJ, Davis CC. 2018 Large-scale digitization of herbarium specimens: development and usage of an automated, high-throughput conveyor system. *Taxon* **67**, 165–178. (doi:10.12705/671.9)
70. Yost JM *et al.* 2018 Digitization protocol for scoring reproductive phenology from herbarium specimens of seed plants. *APPS* **6**. (doi:10.1002/aps3.1022)
71. Williams AC, Goh J, Willis CG, Ellison AM, Brusuelas JH, Davis CC, Law E. 2017 'Deja Vu: characterizing work reliability using task consistency'. In *Proc. of the AAAI Conf. on Human Computation (HCOMP 2017)*, *Quebec City, Canada*. See http://acw.io/pubs/hcomp2017-dejavu.pdf (accessed 14 March 2018)
72. CrowdCurio. See https://crowdcurio.com/ (accessed 14 March 2018)
73. Notes from Nature. See https://www.notesfromnature.org/ (accessed 14 March 2018)
74. BioSpex. See https://biospex.org/ (accessed 14 March 2018)
75. iDigBio. Phenology Working Group. See https://www.idigbio.org/wiki/index.php/Phenology_working_group (accessed 15 February 2018)
76. Meineke EK, Davis CC, Davies TJ. 2018 The unrealized potential of herbaria for global change biology. *Ecol. Monogr*. **88**. (doi:10.1002/ecm.1307)
77. Mohanty SP, Hughes DP, Salathé M. 2016 Using deep learning for image-based plant disease detection. *Front Plant Sci.* **7**, 1–10. (doi:10.3389/fpls.2016.01419)
78. oVert. 2017 Digitization TCN: Collaborative research: oVert: Open exploration of vertebrate diversity in 3D. See https://www.nsf.gov/awardsearch/showAward?AWD_ID=1701714; https://www.floridamuseum.ufl.edu/science/overt/ (accessed 5 January 2018)
79. Grun TB, Nebelsick JH. 2017 Echinoids in 3D: understanding mechanisms that strengthen light-weight skeletons. *Geol. Soc. Am.* **49**, 6. (doi:10.1130/abs/2017AM-293681)
80. eBird. See https://www.ebird.org/home (accessed 5 January 2018)
81. Chatfield-Taylor W, Cole JA. 2017 Living rain gauges: cumulative precipitation explains the emergence schedules of California protoperiodical cicadas. *Ecology* **98**, 2521–2527. (doi:10.1002/ecy.1980)
82. iDigBio. 2013 Developing a centralized digital archive of vouchered animal communication signals. See https://www.idigbio.org/wiki/index.php/Developing_a_Centralized_Digital_Archive_of_Vouchered_Animal_Communication_Signals (accessed 18 January 2018)
83. Macaulay Library. 2018 See https://www.macaulaylibrary.org/ (accessed 17January 2018)
84. Webster M (ed.). 2018 *The extended specimen: emerging frontiers in collections-based ornithological research*. Boca Raton, FL: CRC Press, Taylor and Francis Group.
85. Rabosky ARD, Cox C, Rabosky D, Title PO, Holmes IA, Feldman A, McGuire JA. 2016 Coral snakes predict the evolution of mimicry across New World snakes. *Nat. Commun.* **7**, 11484.
86. Brainerd EL. 2017 Video data and motion analysis in comparative biomechanics research. Presented at the Inaugural Digital Data in Biodiversity Research Conference. See https://www.idigbio.org/wiki/images/7/78/Brainerd_iDigBio2017.pdf (accessed 17 January 2018)
87. Brainerd EL, Baier DB, Gatesy SM, Hedrick TL, Metzger KA, Gilbert SL, Crisco JJ. 2010 X-ray reconstruction of moving morphology (XROMM): precision, accuracy and applications in comparative biomechanics research. *J. Exp. Zool. A Ecol. Genet. Physiol.* **313**, 262–279. (doi:10.1002/jez.589)
88. Webster M, Cicero C, Bates J, Hackett S, Joseph L. 2018 Ornithological collections in the 21st century. In *The extended specimen: emerging frontiers in collections-based ornithological research* (ed. M Webster), pp. 219–232. Boca Raton, FL: CRC Press, Taylor and Francis Group.
89. Peterson AT. 2018 Avifaunal change over three decades in North America detected via integration of specimen and observational data. Presentation given at BCoN Data Integration and Attribution workshop, University of Kansas Biodiversity Institute, Commons, Lawrence, KS, 13–14 February 2018.
90. Peterson AT, Navarro-Sigüenza AG, Martínez-Meyer E, Cuervo-Robayo AP, Berlanga H, Soberón J. 2015 Twentieth century turnover of Mexican endemic avifaunas: landscape change versus climate drivers. *Sci. Adv.* **1**, e1400071. (doi:10.1126/sciadv.1400071)
91. Peterson AT, Navarro-Sigüenza AG, Martínez-Meyer E. 2016 Digital accessible knowledge and well-inventoried sites for birds in Mexico: baseline sites for measuring faunistic change. *PeerJ.* **4**, e2362. (doi:10.7717/peerj.2362)
92. Tobler M, Honorio E, Janovec J, Reynel C. 2007 Implications of collection patterns of botanical specimens on their usefulness for conservation planning: an example of two neotropical plant families (Moraceae and Myristicaceae) in Peru. *Biodivers. Conserv.* **16**, 659–677. (doi:10.1111/ele.12624)
93. Daru BH *et al.* 2017 Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytol.* **217**, 467–960. (doi:10.1111/nph.14855)
94. Meyer C, Weigelt P, Kreft H. 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecol. Lett.* **19**, 992–1006.
95. Pearse WD, Davis CC, Inouye DW, Primack RB, Davies JT. 2017 A statistical estimator for determining the limits of contemporary and historic phenology. *Nat. Ecol. Evol.* **1**, 1876–1882. (doi:10.1038/s41559-017-0350-0)
96. Sikes DS, Copas K, Hirsch T, Longino JT, Schigel D. 2016 On natural history collections, digitized and not: a response to Ferro and Flick. *ZooKeys* **618**, 145–158. (doi:10.3897/zookeys.618.9986)
97. Hill AW, Otegui J, Ariño AH, Guralnick RP. 2010 *GBIF position paper on future directions and recommendations for enhancing fitness-for-use across the GBIF network, version 1.0*. Copenhagen: Global Biodiversity Information Facility, 25 pp. See https://www.gbif.org/document/80623/gbif-

position-paper-on-future-directions-and-recommendations-for-enhancing-fitness-for-use-across-the-gbif-network (accessed 16 March 2018)
98. Cicero C, Spencer C, Bloom D, Guralnick R, Koo M, Otegui J, Russell L, Wieczorek J. 2018 Biodiversity informatics and data quality on a global scale. In *2018 the extended specimen: emerging frontiers in collections-based ornithological research* (ed. M Webster), pp. 201–218. Boca Raton, FL: CRC Press, Taylor and Francis Group.
99. Chapman AD. 2005 Principles and methods of data cleaning – primary species and species-occurrence data, version 1.0. Report for the Global Biodiversity Information Facility. Copenhagen. See https://www.gbif.org/document/80528/principles-and-methods-of-data-cleaning-primary-species-and-species-occurrence-data (accessed 16 March 2018)
100. Smith BE, Johnston MK, Lücking R. 2016 From GenBank to GBIF: phylogeny-based predictive niche modeling tests accuracy of taxonomic identifications in large occurrence data repositories. *PLoS ONE* **11**, e0151232. (doi:10.1371/journal.pone.0151232)
101. Dietrich CH, Dmitriev DA. 2016 Insect phylogenetics in the digital age. *Curr. Opin. Insect. Sci.* **18**, 48–52. (doi:10.1016/j.cois.2016.09.008)
102. MacFadden BJ, Guralnick RP. 2017 Horses in the cloud: big data exploration and mining of fossil and extant *Equus* (Mammalia: Equidae). *Paleobiology* **43**, 1–14. (doi:10.1017/pab.2016.42)
103. Guralnick RP, Zermoglio PF, Wieczorek J, LaFrance R, Bloom D, Russell L. 2016 The importance of digitized biocollections as a source of trait data and a new VertNet resource. *Database* **2016**, 1. (doi:10.1093/database/baw158)
104. Báthori F, Pfliegler WP, Zimmerman C-U, Tartally A. 2017 Online image databases as multi-purpose resources: discovery of a new host ant of *Rickia wasmannii* Cavara (Ascomycota, Laboulbeniales) by screening AntWeb.org. *J. Hymenoptera Res.* **61**, 85–94. (doi:10.3897/jhr.61.20255)
105. Serra-Diaz JM, Enquist BJ, Maitner B, Merow C, Svenning J-C. 2017 Big data of tree species distributions: how big and how good? *Forest Ecosyst.* **4**. (doi:10.1186/s40663-017-0120-0)
106. Zermoglio PF, Guralnick RP, Wieczorek JR. 2016 A standardized reference dataset for vertebrate taxon name resolution. *PLoS ONE* **11**, e0146894. (doi:10.1371/journal.pone.0146894)
107. Eschmeyer WN, Fricke R, van der Laan R (eds). 2018 Catalog of fishes: genera, species, references. See http://researcharchive.calacademy.org/research/ichthyology/catalog/fishcatmain.asp (accessed 21 February 2018)
108. Wilson EO. 2016 *Half-earth: our planet's fight for life*. New York, NY: Liveright Publishing Corporation, a division of W.W. Norton & Company.
109. Roston M. 2015 A guide to digitized natural history collections. *New York Times, Science*. See https://www.nytimes.com/interactive/2015/10/19/science/digitized-museums-guide.html (accessed 16 March 2018).
110. Brockell G. 2017 These three people, and one conveyor belt, are digitizing millions of plant specimens. *Washington Post*. See https://www.washingtonpost.com/video/health-science/these-three-people-and-one-conveyor-belt-are-digitizing-millions-of-plant-specimens/2017/02/08/1719e760-ee15–11e6-a100-fdaaf400369a_video.html?utm_term=8a84c615fc6e (accessed 16 March 2018)
111. Wambsgans EJ. 2017 Endangered insects at the field museum. *Chicago Tribune*. See http://www.chicagotribune.com/news/media/94677154-132.html (accessed 16 March 2018)
112. Lee E. 2016 Museums push to get 'dark data' into light through digitization. *Voice of America News*. See https://www.voanews.com/a/museums-worldwide-push-to-get-dark-data-into-light-through-digitization/3191509.html (accessed 24 January 2018)
113. Paris M. 2014 Canadian nature museum digitizing 3 million specimens. *CBC News*. See http://www.cbc.ca/news/politics/canadian-nature-museum-digitizing-3-million-specimens-1.2482826 (accessed 16 March 2018)
114. Steege H ter, Vaessen RW, Cárdenas-López D, Sabatier D, Antonelli A, de Oliveira SM, Pitman NCA, Jørgensen PM, Salomão RP. 2016 The discovery of the Amazonian tree flora with an updated checklist of all known tree taxa. *Sci. Rep.* **6**, 29549. (doi:10.1038/srep29549)
115. St. Fleur N. 2016 After 300 years of collecting, nearly 12,000 Amazon tree species are found. *The New York Times*. See https://www.nytimes.com/2016/07/13/science/amazon-tree-species-inventory.html (accessed 16 March 2018)
116. Cardoso D *et al.* 2017 Amazon plant diversity revealed by a taxonomically verified species list. *Proc. Natl Acad. Sci. USA* **114**, 10 695–10 700. (doi:10.1073/pnas.1706756114)
117. Allen K. 2016 What can you do with 300,000 dead bees? *Toronto Star*. See https://www.thestar.com/news/insight/2016/01/25/what-can-you-do-with-300000-dead-bees.html (accessed 16 March 2018)
118. Seritan I. 2018 Open House: the inner workings of a museum. *Birding*, February 2018.
119. Hiller AE *et al.* 2017 Mutualism in museums: a model for engaging undergraduates in biodiversity science. *PLoS Biol.* **15**, e2003318. (doi:10.1371/journal.pbio.2003318)
120. Drew JA, Moreau CS, Stiassny MLJ. 2017 Digitization of museum collections holds the potential to enhance researcher diversity. *Nat. Ecol. Evol.* **1**, 1789–1790. (doi:10.1038/s41559-017-0401-6)