Received: 26 October 2018 | Revised: 7 August 2019 | Accepted: 28 August 2019

DOI: 10.1111/ddi.12993

BIODIVERSITY LETTER

Diversity and Distributions WILEY

Born-digital biodiversity data: Millions and billions

Roland Kays[1] | William J. McShea[2] | Martin Wikelski[3,4]

[1]North Carolina Museum of Natural Sciences and North Carolina State University, Raleigh, NC, USA

[2]Smithsonian Conservation Biology Institute, Front Royal, VA, USA

[3]Department of Migration, Max Planck Institute of Animal Behavior, Radolfzell, Germany

[4]Centre for the Advanced Study of Collective Behaviour, University of Konstanz, Radolfzell, Germany

Correspondence
Martin Wikelski, Department of Migration, Germany and Centre for the Advanced Study of Collective Behaviour, Max Planck Institute of Animal Behavior, University of Konstanz, Radolfzell, Germany.
Email: wikelski@ab.mpg.de

Editor: Damaris Zurell

Abstract

Given the dramatic pace of change of our planet, we need rapid collection of environmental data to document how species are coping and to evaluate the impact of our conservation interventions. To address this need, new classes of "born digital" biodiversity records are now being collected and curated many orders of magnitude faster than traditional data. In addition to the millions of citizen science observations of species that have been accumulating over the last decade, the last few years have seen a surge of sensor data, with eMammal's camera trap archive passing 1 million photo-vouchered specimens and Movebank's animal tracking database recently passing 1.5 billion animal locations. Data from digital sensors have other advantages over visual citizen science observation in that the level of survey effort is intrinsically documented and they can preserve digital vouchers that can be used to verify species identity. These novel digital specimens are leading spatial ecology into the era of Big Data and will require a big tent of collaborating organizations to make these databases sustainable and durable. We urge institutions to recognize the future of born-digital records and invest in proper curation and standards so we can make the most of these records to inform management, inspire conservation action and tell natural history stories about life on the planet.

KEYWORDS

animal tracking, big data, biodiversity, camera trapping, conservation, ecological modelling, specimens

Museum specimens have always provided the most basic information about the spatial distribution of life on earth: which species live where and when. These records have formed the basis for our biodiversity range maps, biogeography and conservation planning (Suarez & Tsutsui, 2004). As the pace of global change accelerates, we need more biodiversity data to monitor how species are responding, which are most in need of conservation efforts, and what kinds of impacts these efforts deliver (Dirzo et al., 2014).

A recent paper by Farley, Dawson, Goring, and Williams (2018) discussed ecology's transition into the era of big data and showed exponential increases in biodiversity records in the Global Biodiversity Information Facility (GBIF) and other museum databases. A growing digital archive should put us in a good position to monitor change. However, another recent paper by Malaney and Cook (2018) showed that traditional museums actually are not keeping pace. Mammal specimen collecting in the United States reached its peak around 1990 and has dropped by a factor of three since then, with fewer than 5,000 specimens collected annually in recent years. That this is the situation for North American mammals—one of the world's best surveyed faunas—sheds stark light on what poor resolution incoming specimens will provide to understand changes in our global biodiversity. But what, then, explains the mismatch between the increases in GBIF data and the decreases in actual specimen collection?

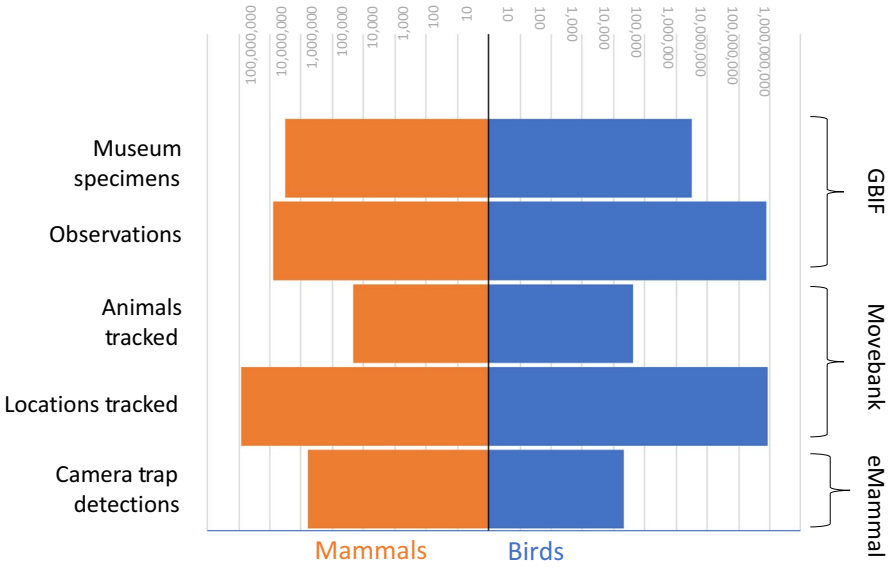

FIGURE 1 Total size of georeferenced datasets available for birds and mammals from GBIF (museum specimens, observations), Movebank (number of individual animals tracked, total locations tracked) and eMammal (camera trap detections). Data available at https://doi.org/10.5061/dryad.b42j56r

1 | BORN DIGITAL BIODIVERSITY

The discrepancy is explained by a new class of biodiversity data that is collected electronically or "born digital". These are not a replacement for physical museum specimens, which are useful in ways that digital collections can never be, including studies of genomic diversity, dietary ecology, disease ecology and morphology, among many other yet undiscovered types of information (Holmes et al., 2016). However, born-digital records are documenting our biodiversity at a faster pace and higher resolution than physical museum specimens ever could. Most of this growth is through human observed data, 98% of GBIF vertebrate records since 2015 are observations (GBIF, 2018), and as of 2019, 94% of all biodiversity records in GBIF were observations. The volume of these observations has clearly led to new insight, enabled in part by sophisticated data filtering algorithms (Kelling, Yu, Gerbracht, & Wong, 2011), but the accuracy of these observations is typically impossible to check since most do not have any record that can be verified (i.e. no voucher specimen retained as a reference); indeed <1% have associated media that could function as a photograph or acoustic voucher. Furthermore, Bayraktarov et al. (2019) question whether the big unstructured biodiversity data provided by non-standardized surveys really mean more knowledge. Approaches that do not document details of sampling effort, or give incomplete species records, will be of dubious value for modelling efforts to establish predictive relationships between species and environmental conditions (Bayraktarov et al., 2019; Steger, Butt, & Hooten, 2017). Fortunately, two sensor-driven types of born-digital biodiversity data, camera traps and animal tracking devices, are now maturing and coalescing to provide verifiable big data with well-documented sampling protocols and survey effort (Kays, Crofoot, Jetz, & Wikelski, 2015; Steenweg et al., 2017). The scale of data collected by these sensors has rapidly caught up with museums and citizen observations (Figure 1). While not a solution for all groups, existing data represent a diversity of bird and mammal groups, around the world, including species of conservation concern (Figure 2).

As a photo-vouchered spatial record of biodiversity, camera traps offer a direct parallel to the museum mammal specimen because the identity of the species can be verified in the photograph, potentially even automatically through artificial intelligence (He et al., 2016). Although not all species can be visually distinguished (Potter, Brady, & Murphy, 2018), camera traps are useful for most medium or large terrestrial mammals and have recently proven effective for small mammals and canopy fauna (Bowler, Tobler, Endress, Gilmore, & Anderson, 2016; McCleery et al., 2014). Camera traps also have the advantage of clearly recording sampling effort (where they are run and for how long), which is typically not known for museum collections or citizen science observations. Since building eMammal as a repository for camera trap photographs at the Smithsonian in 2012, we have seen steady growth of records and by 2019 have > 1 million georeferenced, vouchered animal records (Figure 1). To put this in perspective, the world's largest physical mammal collection, also at the Smithsonian, has just under 600,000 georeferenced mammal records spanning 180 years, and the second-largest mammal collection (Museum of Southwestern Biology) has about half that (Dunnum et al., 2018). Furthermore, eMammal probably represents a relatively small amount of the camera trap data collected in the last decade. A new collaborative camera trapping project called Wildlife Insights will provide the artificial intelligence and automated analytical tools to process and analyse big data efficiently, thereby bringing together even more born digital data from around the world for effective and more timely monitoring of animal life on Earth.

Modern GPS technology has empowered the animal tracking field to grow even faster. For example, the Movebank animal tracking database we established at the Max Planck Institute of Animal Behaviour in 2009 has over 1.5 billion georeferenced animal records (Figure 1). Animal tracking data are inherently autocorrelated and so do not function as statistically independent occurrences in spatial models as museum specimens typically do. However, this more detailed

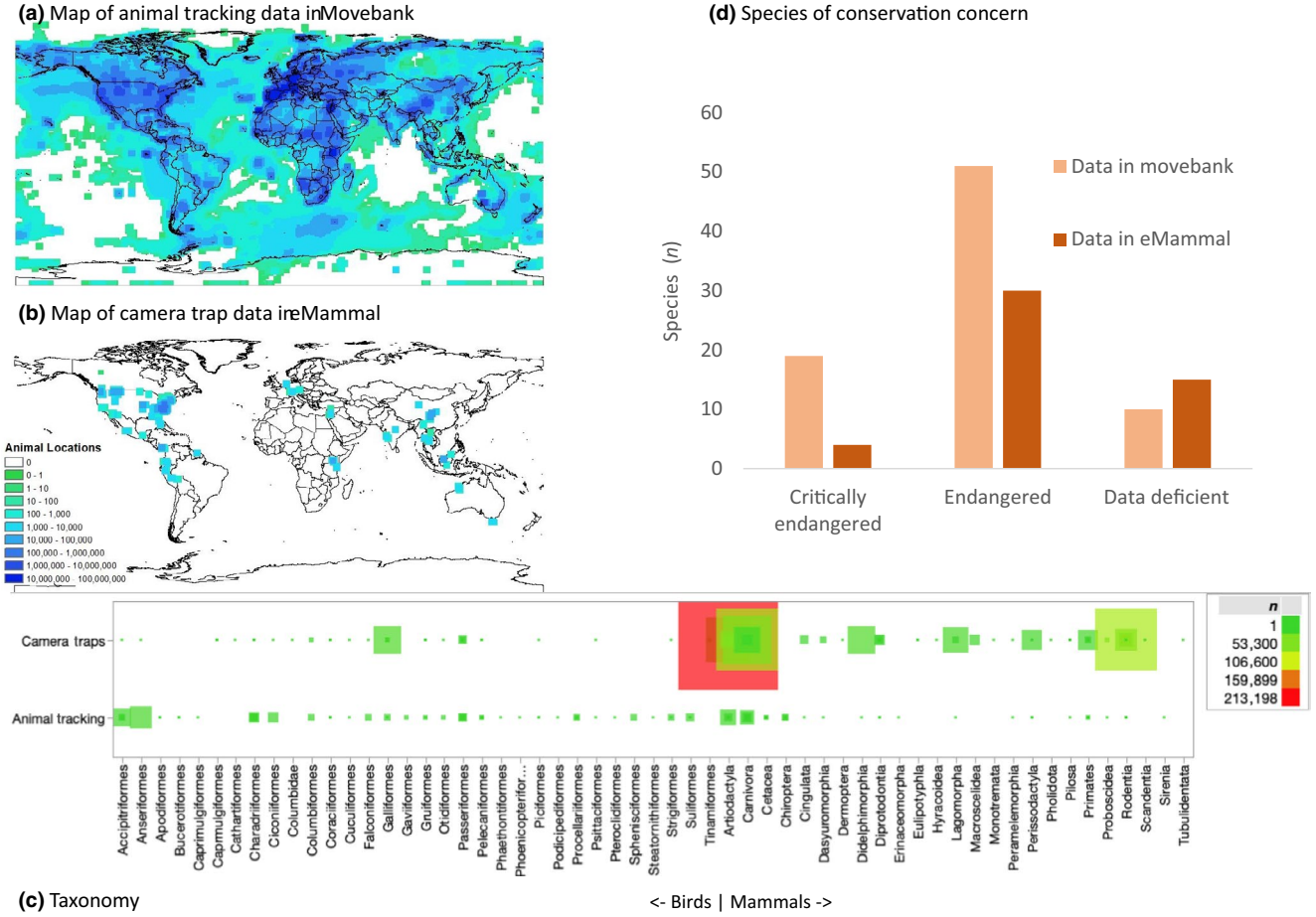

FIGURE 2 Geographic and Taxonomic scope of born digital data from animal tracking (Movebank) and camera trapping (eMammal) databases. Maps show the number of tracking locations (a, publicly available animal tracking data) or number of animals detected (b, camera traps). The colour scale from B also applies to A. The colour and size of squares in graph (c) show the number of detections (camera traps) or number of individuals tracked (animal tracking) from orders and families of birds and mammals while (d) shows the number of species with born digital data that are classified as Critically Endangered, Endangered or Data Deficient by the IUCN Red Data List. eMammal has data from 632 species of birds and mammals from 121 families and 41 orders while Movebank has tracking data from 805 bird and mammal species representing 138 families and 45 orders. Movebank also has data from Amphibia, Cephalaspidomorphi, Chondrichthyes, Chondrostei, Insecta, Plantae, Reptilia and Teleostei that are not illustrated in these graphs. Taxonomic data are available at https://doi.org/10.5061/dryad.b42j56r

picture of how animals use space can help address a wealth of questions about habitat suitability, ecological interactions and response to human disturbance (Kays et al., 2015). Additionally, new data fusion statistical techniques are providing the framework to combine the fine-scale inference of animal tracking data with the larger scale reference of other biodiversity data (Pacifici et al., 2017). As tracking tag technology miniaturizes, we can track smaller and smaller species (Kays et al., 2015). For example, the new ICARUS antenna that was recently mounted on the International Space Station allows the global tracking of 5 g GPS transmitters suitable for tracking 100 g birds with GPS accuracy and near global data readout (Wikelski et al., 2007).

2 | USES FOR BORN-DIGITAL RECORDS

The collection of Born Digital data in the first place is motivated primarily by spatial ecology, with animal tracking usually considering questions at the level of individuals or populations and camera trapping assessing populations or communities. The improvement in data management and decrease in cost of sensors have enabled larger-scale studies (Kays et al., 2015; Steenweg et al., 2017) and the creation of long-term monitoring projects (Rovero & Ahumada, 2017). Data sharing across projects, supported by appropriate cyber-infrastructure, has also allowed scientists to ask basic questions at a global scale, for example, how do humans affect animal movement (Tucker et al., 2018). Given their high spatial accuracy, large volume and explicit measures of effort, these Born Digital data are expected to play a big part in the estimation of Essential Biodiversity Variables used to assess progress towards the objectives of the Convention on Biological Diversity (CBD) and Sustainable Development Goals (SDGs) (Jetz et al., 2019).

In addition to these empirical uses of Born Digital data, we also see great potential to use the images and stories of these animals to help connect people with nature and inspire them to contribute to

the conservation values that motivate global efforts like CBD and SDGs. The stories of individually tracked animals, like Pluie the wolf (Yellowstone to Yukon, Locke & Heuer, 2015) and Alice the moose (Algonquin to Adirondacks, Braszak, 2017), have already inspired large scale conservation efforts. High-resolution tracking data of modern studies are now engaging millions of people to follow the migrations of animals in real time through websites or apps (e.g. Animal Tracker). Amazing animal pictures from camera traps are now shared via social media, websites or books (Kays, 2016) as standard public engagement tools for scientists and conservation organizations. Born Digital data collection can also be integrated into citizen science projects, providing verifiable data to scientists and unique experiences for volunteers that can have positive conservation impacts. For example, Roetman et al. (2017) showed that volunteers who used GPS units to track their pet cats movement changed their behaviour to limit their pet's hunting of native prey. Similarly, Forrester et al. (2016) showed that volunteers who helped run camera traps for research became advocates for conservation of their local fauna.

3 | DIGITAL BIODIVERSITY INFRASTRUCTURE SUPPORT

As Farley et al. (2018) point out, the transition of a field into the era of big data requires the consecutive development of technology to collect the data, statistical tools to analyse them and cyberinfrastructure to manage them. While the development and popularization of digital camera traps and miniaturized GPS trackers in the last decade started the boon in born-digital biodiversity data, making efficient use of this information has also required new cyberinfrastructure to handle this flood of data. These big ecology databases need to be online continuously to enable live data streams coming in and collaborative data sharing going out. For example, Movebank now has consecutive live feeds from ca. 5,000 animals via GSM or satellite networks delivering approximately 1 million animal locations per day, while eMammal has ca. 10K camera trap detections uploaded per week. While most statistical analyses are performed locally by scientists, there is also a need for web-based analytics to enable real-time monitoring and also make data available to users as diverse as land managers or school children (Schuttler et al., 2018).

The cyberinfrastructure and data curation that makes big data ecology possible are expensive. Not only do these require extensive bandwidth and server space, but also web interfaces and analytical tools. A database is never “done” but needs continual support to pay for the never-ending updates that maintain security and connectivity, not to mention upgrades to support new user needs and data streams. The natural history museums that hold our physical specimens are one logical home for these cybertools, but broader collaboration is needed across government organizations, NGOs, universities and research institutes to bear the annual costs of supporting born digital big data biodiversity. End users should also recognize the value of born digital data and tools to their work and expect some payment for services to be part of future funding models. Charging for data access would be against the spirit of open data and discourage wide use of these resources, particularly for developing countries that host much of the planet´s biodiversity. We believe that data should be freely ingested and freely provided in standard format. However, we suggest it is appropriate to charge for premium services such as streamlined ingestion of very large data sets (e.g. >1 hz sensor streams), more complicated derived data products or feature-heavy analytic protocols that would not only help sustain this cyberinfrastructure, but also widen the potential audience of users for these data.

Natural history museums were created as institutions to protect physical specimens so they are available to researchers for perpetuity and to use the objects and science stories to engage and educate a broad audience through exhibits and programming. Born-digital biodiversity data has the same potential for research and engagement value, but instead of shelving and taxidermists, we need to invest in servers, programmers and apps if we are to make them work as long-term records of planetary change, and inspiration for people to care about the natural world.

ACKNOWLEDGEMENTS

Big thanks to the larger teams that make eMammal and Movebank possible. We thank the Max Planck Institute of Animal Behaviour, The Smithsonian Institute, the North Carolina Museum of Natural Sciences, NASA, and the National Science Foundation for supporting Movebank and eMammal.

ORCID

Roland Kays https://orcid.org/0000-0002-2947-6665

REFERENCES

Bayraktarov, E., Ehmke, G., O'Connor, J., Burns, E. L., Nguyen, H. A., McRae, L., … Lindenmayer, D. B. (2019). Do big unstructured biodiversity data mean more knowledge? *Frontiers in Ecology and Evolution*, *6*, 239. https://doi.org/10.3389/fevo.2018.00239

Bowler, M. T., Tobler, M. W., Endress, B. A., Gilmore, M. P., & Anderson, M. J. (2016). Estimating mammalian species richness and occupancy in tropical forest canopies with arboreal camera traps. *Remote Sensing in Ecology and Conservation*, *3*, 1–12. https://doi.org/10.1002/rse2.35

Braszak, P. (2017). *Social Movement Theory and Transboundary Conservation in Eastern North America: A Case Study of the Algonquin to Adirondacks Collaborative (University of Toronto)*. Retrieved from https://tspace.library.utoronto.ca/bitstream/1807/79218/3/Braszak_Patrick_201711_MA_thesis.pdf

Dirzo, R., Young, H. S., Galetti, M., Ceballos, G., Isaac, N. J. B., & Collen, B. (2014). Defaunation in the anthropocene. *Science*, *345*(6195), 401–406. https://doi.org/10.1126/science.1251817

Dunnum, J. L., McLean, B. S., Dowler, R. C., Bradley, J. E., Bradley, R. D., Carraway, L. N., … Velazco, P. M. (2018). Mammal collections of the Western Hemisphere: A survey and directory of collections. *Journal of Mammalogy*, *99*(6), 1307–1322. https://doi.org/10.1093/jmammal/gyy151

Farley, S. S., Dawson, A., Goring, S. J., & Williams, J. W. (2018). Situating ecology as a big-data science: current advances, challenges, and solutions. *BioScience*, *68*(8), 563–576. https://doi.org/10.1093/biosci/biy068

Forrester, T. D., Baker, M., Costello, R., Kays, R., Parsons, A. W., & McShea, W. J. (2016). Creating advocates for mammal conservation through citizen science. *Biological Conservation*, *208*, 98–105. https://doi.org/10.1016/j.biocon.2016.06.025

GBIF (2018) *GBIF records come from an online search of their database in August 2018*. (n.d.).

He, Z., Kays, R., Zhang, Z., Ning, G., Huang, C., Han, T. X., … McShea, W. (2016). Visual informatics tools for supporting large-scale collaborative wildlife monitoring with citizen scientists. *IEE Circuits and Systems Magazine*, *16*, 73–86. https://doi.org/10.1109/MCAS.2015.2510200

Holmes, M. W., Hammond, T. T., Wogan, G. O. U., Walsh, R. E., LaBarbera, K., Wommack, E. A., … Nachman, M. W. (2016). Natural history collections as windows on evolutionary processes. *Molecular Ecology*, *25*(4), 864–881. https://doi.org/10.1111/mec.13529

Jetz, W., McGeoch, M. A., Guralnick, R., Ferrier, S., Beck, J., Costello, M. J., … Turak, E. (2019). Essential biodiversity variables for mapping and monitoring species populations. *Nature Ecology & Evolution*, *3*, 539–551. https://doi.org/10.1038/s41559-019-0826-1

Kays, R. (2016). *Candid creatures: How camera traps reveal the mysteries of nature*. Baltimore, MD: Johns Hopkins University Press.

Kays, R., Crofoot, M. C., Jetz, W., & Wikelski, M. (2015). Terrestrial animal tracking as an eye on life and planet. *Science*, *348*(6240), aaa2478–aaa2478. https://doi.org/10.1126/science.aaa2478

Kelling, S., Yu, J., Gerbracht, J., & Wong, W.-K. (2011). Emergent filters: Automated data verification in a large-scale citizen science project. *IEEE Seventh International Conference on E-Science Workshops*, *2011*, 20–27. https://doi.org/10.1109/eScienceW.2011.13

Locke, H., & Heuer, K. (2015). *Yellowstone to yukon: global conservation innovations through the years*. In Protecting the Wild (pp. 120–130).

Malaney, J. L., & Cook, J. A. (2018). A perfect storm for mammalogy: Declining sample availability in a period of rapid environmental degradation. *Journal of Mammalogy*, *99*(4), 773–788. https://doi.org/10.1093/jmammal/gyy082

MCCleery, R. A., Zweig, C. L., Desa, M. A., Hunt, R., Kitchens, W. M., & Percival, H. F. (2014). A novel method for camera-trapping small mammals. *Wildlife Society Bulletin*, *38*(4), 887–891. https://doi.org/10.1002/wsb.447

Pacifici, K., Reich, B., Miller, D., Gardner, B., Glenn, S., Singh, S., … Collazo, J. (2017). Integrating multiple data sources in species distribution modeling: A framework for data fusion. *Ecology*, *98*(3), 840–850. https://doi.org/10.13140/RG.2.1.3702.8566

Potter, L. C., Brady, C. J., & Murphy, B. P. (2018). Accuracy of identifications of mammal species from camera trap images: A northern Australian case study. *Austral Ecology*, *44*, 473–483. https://doi.org/10.1111/aec.12681

Roetman, P., Tindle, H., Litchfield, C., Chiera, B., Quinton, G., Kikillus, H., & Kays, R. (2017). *Cat tracker South Australia: Understanding pet cats through citizen science*.

Rovero, F., & Ahumada, J. (2017). The tropical ecology, assessment and monitoring (TEAM) network: An early warning system for tropical rain forests. *Science of the Total Environment*, *574*, 914–923. https://doi.org/10.1016/J.SCITOTENV.2016.09.146

Schuttler, S. G., Sears, R. S., Orendain, I., Khot, R., Rubenstein, D., Rubenstein, N., … Kays, R. (2018). Citizen science in schools: Students collect valuable mammal data for science, conservation, and community engagement. *BioScience*, *69*, 69–79. https://doi.org/10.1093/biosci/biy141

Steenweg, R., Hebblewhite, M., Kays, R., Ahumada, J., Fisher, J. T., Burton, C., … Rich, L. N. (2017). Scaling up camera traps: Monitoring the planet's biodiversity with networks of remote sensors. *Frontiers in Ecology and the Environment*, *15*, 26–34. https://doi.org/10.1002/fee.1448

Steger, C., Butt, B., & Hooten, M. B. (2017). Safari Science: Assessing the reliability of citizen science data for wildlife surveys. *Journal of Applied Ecology*, *54*(6), 2053–2062. https://doi.org/10.1111/1365-2664.12921

Suarez, A. V., & Tsutsui, N. D. (2004). The value of museum collections for research and society. *BioScience*, *54*, 66–74. https://doi.org/10.1641/0006-3568(2004)054[0066:TVOMCF]2.0.CO;2

Tucker, M. A., Böhning-Gaese, K., Fagan, W. F., Fryxell, J. M., Van Moorter, B., Alberts, S. C., … Mueller, T. (2018). Moving in the Anthropocene: Global reductions in terrestrial mammalian movements. *Science*, *359*(6374), 466–469. https://doi.org/10.1126/science.aam9712

Wikelski, M., Kays, R., Kasdin, J., Thorup, K., Smith, J. A., Cochran, W. W., & Swenson, G. W. Jr (2007). Going wild: What a global small-animal tracking system could do for experimental biologists. *Journal of Experimental Biology*, *210*, 181–186. https://doi.org/10.1242/jeb.02629

BIOSKETCHES

Roland Kays is the Head of the Biodiversity Lab at the NC Museum of Natural Sciences, a Research Professor in the department of Forestry and Environmental Resources at NC State University, and co-pi of Movebank and eMammal.

William J. McShea is a Wildlife Ecologist at the Smithsonian Conservation Biology Institute and co-pi of eMammal.

Martin Wikelski is the director of the Department of Migration, Max Planck Institute of Animal Behavior, Radolfzell, Germany, Board Member of the Centre for the Advanced Study of Collective Behaviour, University of Konstanz, Konstanz, and co-pi of Movebank.

Author contributions: All authors shared in conceiving of the ideas and in the writing, R.K. analysed the data.

How to cite this article: Kays R, McShea WJ, Wikelski M. Born-digital biodiversity data: Millions and billions. *Divers Distrib*. 2019;00:1–5. https://doi.org/10.1111/ddi.12993