BIOLOGY LETTERS

rsbl.royalsocietypublishing.org

Research

Cite this article: Marshall CR *et al*. 2018 Quantifying the dark data in museum fossil collections as palaeontology undergoes a second digital revolution. *Biol. Lett.* **14**: 20180431.
http://dx.doi.org/10.1098/rsbl.2018.0431

Received: 13 June 2018
Accepted: 6 August 2018

Subject Areas:
palaeontology

Keywords:
digitization, dark data, museum collections, iDigBio

Author for correspondence:
C. R. Marshall
e-mail: crmarshall@berkeley.edu

THE ROYAL SOCIETY PUBLISHING

Palaeontology

Quantifying the dark data in museum fossil collections as palaeontology undergoes a second digital revolution

C. R. Marshall[1,2], S. Finnegan[1,2], E. C. Clites[2], P. A. Holroyd[2], N. Bonuso[3], C. Cortez[4], E. Davis[5,6], G. P. Dietl[7,8], P. S. Druckenmiller[9], R. C. Eng[10], C. Garcia[11], K. Estes-Smargiassi[12], A. Hendy[12], K. A. Hollis[13], H. Little[13], E. A. Nesbitt[10], P. Roopnarine[11], L. Skibinski[7], J. Vendetti[12] and L. D. White[2]

[1]Department of Integrative Biology, University of California, 3040 Valley Life Sciences Building, Berkeley, CA 94720-3140, USA
[2]University of California Museum of Paleontology, University of California, 1101 Valley Life Sciences Building, Berkeley, CA 94720-4780, USA
[3]Department of Geological Sciences, California State University, Fullerton, CA 92834, USA
[4]John D. Cooper Archaeological and Paleontological Center, Santa Ana, CA 92701-6427, USA
[5]Department of Earth Sciences, University of Oregon, Eugene, OR 97403-1272, USA
[6]University of Oregon Museum of Natural and Cultural History, 1680 E. 15th Avenue, Eugene, OR 97403-1224, USA
[7]Paleontological Research Institution, 1259 Trumansburg Road, Ithaca, NY 14850, USA
[8]Department of Earth and Atmospheric Sciences, Cornell University, 112 Hollister Drive, Ithaca, NY 14853, USA
[9]University of Alaska Museum and Department of Geosciences, University of Alaska Fairbanks, 1962 Yukon Drive, Fairbanks, AK 99775, USA
[10]Burke Museum of Natural History and Culture, University of Washington, Box 353010, Seattle, WA 98195-3010, USA
[11]California Academy of Sciences, 55 Music Concourse Drive, San Francisco, CA 94118, USA
[12]Natural History Museum of Los Angeles County, 900 Exposition Boulevard, Los Angeles, CA 90007, USA
[13]Department of Paleobiology, National Museum of Natural History, Smithsonian Institution, PO Box 37012, Washington, DC 20013, USA

CRM, 0000-0001-7832-0950; SF, 0000-0002-6175-6173; ECC, 0000-0002-5264-8154; PAH, 0000-0003-1292-6356; ED, 0000-0002-0918-5852; GPD, 0000-0003-1571-0868; AH, 0000-0002-9818-1158; JV, 0000-0003-2260-7762

Large-scale analysis of the fossil record requires aggregation of palaeontological data from individual fossil localities. Prior to computers, these synoptic datasets were compiled by hand, a laborious undertaking that took years of effort and forced palaeontologists to make difficult choices about what types of data to tabulate. The advent of desktop computers ushered in palaeontology's first digital revolution—online literature-based databases, such as the Paleobiology Database (PBDB). However, the published literature represents only a small proportion of the palaeontological data housed in museum collections. Although this issue has long been appreciated, the magnitude, and thus potential significance, of these so-called 'dark data' has been difficult to determine. Here, in the early phases of a second digital revolution in palaeontology—the digitization of museum collections—we provide an estimate of the magnitude of palaeontology's dark data. Digitization of our nine institutions' holdings of Cenozoic marine invertebrate collections from California, Oregon and Washington in the USA reveals that they represent 23 times the number of unique localities than are currently available in the PBDB. These data, and the vast quantity of similarly untapped dark data in other museum collections, will, when digitally mobilized, enhance palaeontologists' ability to make inferences about the patterns and processes of past evolutionary and ecological changes.

1. Palaeontology's first digital revolution

Large-scale analysis of evolutionary and ecological patterns in the fossil record [1] was pioneered by single investigators, from Phillips [2] to those derived

from the compendia of Sepkoski [3,4], each of which took years to compile by hand. Multi-authored data compilations were also undertaken, including by Hallam [5] and Benton [6]. Typically, the burden of compiling the data was so great that valuable ancillary data were not tabulated, making it virtually impossible to extend the depth or sophistication of analyses performed with these databases. For example, Sepkoski's compendia tabulate first and last occurrences only, with no geographical data, nor data on the richness of the fossil record for each taxon, nor their taphonomy, abundance, palaeoecology, etc., nor, for extant taxa, the age of the youngest known fossil occurrences.

Stymied by these limitations, the palaeobiology community has undertaken several initiatives to digitally aggregate data from the primary literature to enable rapid large-scale synthetic analyses of the fossil record. This first digital revolution in palaeobiology has resulted in several still growing databases, for example, the New and Old Worlds (NOW) Database of fossil mammals (http://www.helsinki.fi/science/now/), the Neotoma Paleoecological Database consortium (http://www.neotomadb.org/), among others [7]. The temporally, geographically and taxonomically most comprehensive is the literature-based Paleobiology Database (PBDB) (http://paleobiodb.org/), although it was initially compiled to answer questions that only required the data to be taxonomically and temporally representative rather than comprehensive—presently for most taxa, the PBDB is still not comprehensive. Nonetheless, the database currently includes 410 contributors, and information on 194 000 collection sites, 371 000 taxa and 1.37 million fossil occurrences. It has enabled 317 publications, and has motivated an annual international graduate student summer workshop in palaeontological data analysis (http://www.analytical.palaeobiology.de/).

2. Palaeontology's second digital revolution

The published literature, although rich, documents only a fraction of the fossils housed in the world's museums [8]. To date, it has been almost impossible to estimate the quantity of these additional dark data [9], but now a second digital revolution is under way in palaeontology—the digital aggregation of these unpublished and largely inaccessible fossil collections and their metadata. Within the USA, this work is being led by the National Science Foundation's (NSF) Division of Biological Infrastructure (DBI) program for Advancing Digitization of Biodiversity Collections (ADBC), currently via 20 thematic collections networks (TCNs) (http://www.idigbio.org/content/thematic-collections-networks). Among these are four palaeontological TCNs: (i) The Cretaceous World (fossils from the Western Interior Seaway), (ii) Fossil Insect Collaborative, (iii) PALEONICHES (marine faunas of the Ordovician, Pennsylvanian and Neogene) and (iv) EPICC (Eastern Pacific invertebrate Cenozoic communities of marine fossils from Alaska to Chile), which is the focus of the authors of this paper. Typically, the metadata being captured by each fossil TCN includes stratigraphic, geochronologic and georeferenced locality data for each collection site, and the imaging of representative specimens.

Given the dispersed distribution of museum collections, fossil or otherwise, a critical component of this second digitization revolution is the development of a one-stop point of online access to the digital data, the Integrated Digitized Biocollections database [10] (iDigBio) (http://www.idigbio.org/). Importantly, iDigBio is promoting, sharing and coordinating best practices and protocols so that all museums can take maximal advantage of this effort to digitize museum collections, and to ensure the continuity and thus enduring value of the data. For palaeontological data new tools such as the Enhancing Paleontological and Neontological Data Discovery API (ePANDDA) (http://epandda.org/) that will link museum data with the PBDB and the Macrostrat (http://macrostrat.org/) map database, among other databases, will enable the use of this vast volume of previously untapped data to inform big data science in palaeobiology [8].

Mobilization of these museum records will be of enormous scientific value, enabling, for example, more precise estimates of geographic and stratigraphic ranges, improved knowledge of the distribution and partitioning of taxa and faunal associations across environmental gradients, enhanced ability to characterize morphological clines and identify ecophenotypic effects, and more opportunities to identify specimens suitable for morphological and stable isotopic analyses.

3. A quantification of the amount of dark data in fossil collections

By drawing on a subset of the specimen data currently being digitized by the EPICC TCN (http://epicc.berkeley.edu/), those from just California, Oregon and Washington, we have been able to provide a measure of just how much more data are housed in museum collections than are available in the PBDB. The museum collections data we analysed are relatively broad taxonomically (fossil invertebrates), represent a large slice of geological time (the Cenozoic Era, from 66 million years ago to present) and are from a geographically well-defined and relatively fossiliferous region (the west coast of the lower 48 states of the USA). Our analysis reveals that nine of our TCN's collections have fossils from approximately 23 times the number of marine fossil localities than are currently entered in the PBDB (figure 1). This suggests that globally perhaps only 3–4% of recorded fossil localities are currently accounted for in the PBDB. Given that the standard error of any parameter estimate (e.g. means), or the uncertainty in the slope of a line of best fit, is approximately proportional to the square root of the number of data points analysed, this result indicates that the EPICC TCN museum data (when fully mobilized) will offer an approximately fivefold increase in the precision of such estimates over those made with the data currently available in the PBDB.

4. Significance of palaeontology's second digital revolution

In this age of rapid global change, palaeontological knowledge of past evolutionary and ecological changes and their causes and consequences is especially relevant to understanding life and its future on Earth [12]. However, palaeontology's second digital revolution is still in its infancy with only a tiny proportion of museums' dark data being digitally mobilized, and only a subset of higher taxa and geographical regions currently being targeted. We advocate that efforts to bring these dark data to light be continued and expanded. Funding such efforts would represent a relatively small proportion of the resources already put towards maintenance of these museum

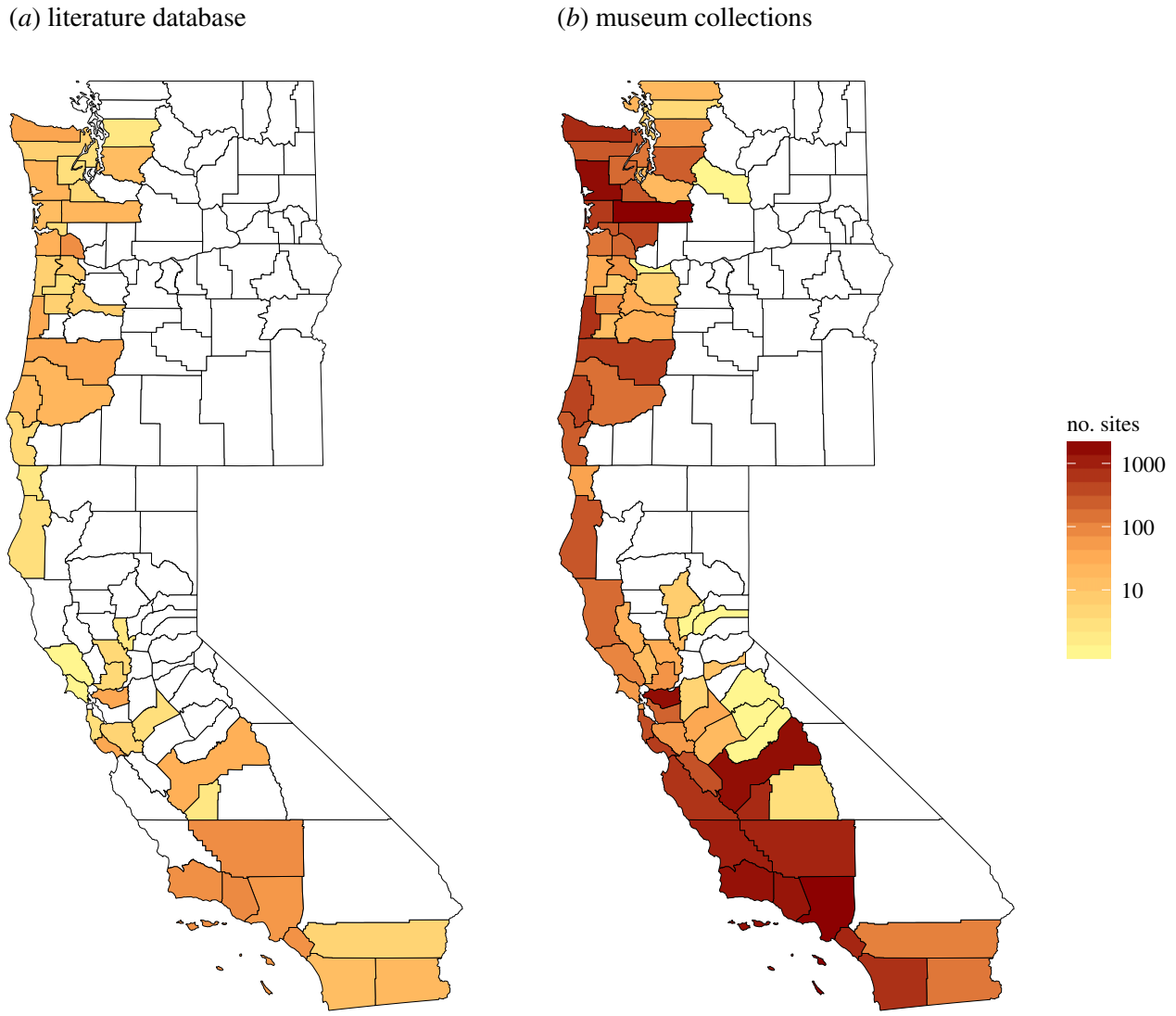

Figure 1. Visualization of the 23-fold increase in digitally accessible Cenozoic marine invertebrate palaeontological collection sites (26 059) from museum collections compared with the number of collection sites (1139) from literature data currently entered into the PBDB (https://paleobiodb.org/) for California, Oregon and Washington. (*a*) Number of sites per county currently included in the PBDB (https://paleobiodb.org/); (*b*) number of sites per county now digitally mobilized across nine institutions of the EPICC TCN (https://epicc.berkeley.edu/). The number of sites per county for each map are provided in the Supplemental_Data.csv file deposited in the Dryad data repository (doi:10.5061/dryad.j0r8127) [11].

collections, and thus, funds allocated in this area would yield an excellent return on total investment. Likewise, we must continue to fund the infrastructure that supports socially and scientifically vital museum collections, reversing a current trend of divestment by many governing organizations [9].

Data accessibility. The raw data tabulated from the EPICC TCN are in the process of being uploaded to iDigBio by each institution. Tabulated data for figure 1 can be found in the Dryad data package: http://dx.doi.org/10.5061/dryad.j0r8127 [11] in the file Supplemental_Data.csv.

Authors' contributions. All authors participated in the discussions that led to this paper, were involved in the collection of the data, contributed to interpreting the results, writing the manuscript, are accountable for all aspects of the work and agree to its publication. E.C.C. oversaw the gathering of the data. S.F. drafted the figure. C.R.M. led the drafting of the manuscript. This is Paleobiology Database contribution number 317.

Competing interests. We have no competing interests.

Funding. This work was funded by NSF ADBC awards 1503678, 1503628, 1503611, 1503065, 1503613, 1503545, 1502500, and NSF CSBR awards 1349430, 1561429, 1561759 and 1203600.

Acknowledgements. We thank all the graduate and undergraduate students and volunteers who have worked on the EPICC TCN project.

References

1. Sepkoski D. 2009 *Rereading the fossil record: the growth of paleobiology as an evolutionary discipline*. Chicago, IL: University Chicago Press.
2. Phillips J. 1860 *Life on the earth: its origin and succession*. Cambridge, UK: Macmillan and Company.
3. Sepkoski Jr JJ. 1992 A compendium of fossil marine animal families. *Milwaukee Public Museum Contrib. Biol. Geol.* **83**, 1–156.
4. Sepkoski Jr JJ. 2002 A compendium of fossil marine animal genera. *Bull. Am. Paleontol.* **363**, 1–560.
5. Hallam A (ed.). 1977 *Patterns of evolution as illustrated by the fossil record*. Amsterdam, The Netherlands: Elsevier Scientific Publishing Company.
6. Benton MJ (ed.). 1993 *The fossil record 2*. London, UK: Chapman and Hall.

7. Uhen MD *et al.* 2013 From card catalogs to computers: databases in vertebrate paleontology. *J. Vertebr. Paleontol.* **33**, 13–28. (doi:10.1080/02724634.2012.716114)
8. Allmon WA, Dietl GP, Hendricks JR, Ross RM. 2018 Bridging the two fossil records: paleontology's 'big data' future resides in museum collections. In *Museums at the forefront of the history and philosophy of geology: history made, history in the making* (eds GD Rosenberg, R Clary). Special paper 535. Boulder, CO: Geological Society of America.
9. Rogers N. 2016 Museum drawers go digital. *Science* **352**, 762–763.
10. Page LM, MacFadden BJ, Fortes JA, Soltis PS, Riccardi G. 2015 Digitization of biodiversity collections reveals biggest data on biodiversity. *Bioscience* **65**, 841–842.
11. Marshall CR *et al.* 2018 Data from: Quantifying the dark data in museum fossil collections as palaeontology undergoes a second digital revolution. Dryad Digital Repository. (doi:10.5061/dryad.j0r8127)
12. Barnosky AD *et al.* 2017 Merging paleobiology with conservation biology to guide the future of terrestrial ecosystems. *Science* **355**, eaah4787. (doi:10.1126/science.aah4787)