Bridging the two fossil records: Paleontology's "big data" future resides in museum collections

Chapter · May 2018

DOI: 10.1130/2018.2535(03)

CITATIONS
11

READS
970

4 authors, including:

Warren D. Allmon
Paleontological Research Institution
194 PUBLICATIONS **2,739** CITATIONS

SEE PROFILE

Jonathan R. Hendricks
Paleontological Research Institution
47 PUBLICATIONS **969** CITATIONS

SEE PROFILE

Robert M. Ross
Paleontological Research Institution
86 PUBLICATIONS **482** CITATIONS

SEE PROFILE

The Geological Society of America
Special Paper 535

Bridging the two fossil records: Paleontology's "big data" future resides in museum collections

Warren D. Allmon*,†
Gregory P. Dietl*
Jonathan R. Hendricks*
Robert M. Ross
Paleontological Research Institution, 1259 Trumansburg Road, Ithaca, New York 14850, USA, and
Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, New York 14853, USA

ABSTRACT

There are two fossil records: the *physical fossil record,* which consists of specimens, and the *abstracted fossil record,* which is made up of data derived from those specimens. Museum collections are the conduit between these two fossil records. Over the past several decades, the abstracted fossil record has provided many important insights about the major features of life's history, but it has relied mostly on limited types of data (primarily taxonomic occurrence data) derived from ultimately finite literature sources. In contrast, specimen collections and modern tools for digitizing information about them present an opportunity to transform paleobiology into a "big data" science. Digitally capturing non-traditional (e.g., paleoecological, taphonomic, geochemical, and morphological) data from millions of specimens in museum collections and then integrating them with other unique big data resources has the potential to lead to the most important paleontological discoveries of the twenty-first century.

What we know about the past record relied heavily on museum collections—the cumulation of centuries of investigation of the fossil record. The sample of past biodiversity will accumulate only with continued exploration of the fossil record … and restudy of existing collections….
—J. Sepkoski (1992, p. 80)

*Authors contributed equally and are listed alphabetically.
†Corresponding author: wda1@cornell.edu.

Allmon, W.D., Dietl, G.P., Hendricks, J.R., and Ross, R.M., 2018, Bridging the two fossil records: Paleontology's "big data" future resides in museum collections, *in* Rosenberg, G.D., and Clary, R.M., eds., Museums at the Forefront of the History and Philosophy of Geology: History Made, History in the Making: Geological Society of America Special Paper 535, https://doi.org/10.1130/2018.2535(03).

INTRODUCTION: THE TWO FOSSIL RECORDS

There are two fossil records, and the data-driven science of paleontology balances upon both, like a beam that spans two pillars. The first fossil record—which we call the *physical fossil record*—consists of material objects: discovered specimens observed in the field or (more frequently) residing in collections, as well as specimens remaining undiscovered in nature. Yet fossils are not just things we can observe, they are also objects rich with biological, temporal, and geographic context. Because each fossil provides a record of the *consequences* of past biological and physical processes—rather than direct observation of past organisms, environments, and processes themselves—paleontology, like all historical sciences, is to some degree a science of inference and abstraction. Interpreting a fossil as anything other than a rock requires inference and abstraction based on context and comparison with modern observable analogs. This contextual and comparative information forms the second pillar of the fossil record, which we call the *abstracted fossil record*.[1]

Paleobiologists frequently divide themselves between two camps: those who tout their "specimen-based research," and those who identify as "quantitative" or "analytical" paleobiologists, often spending more time with computer databases than specimen collections. Both approaches, of course, rely upon abstracted information derived from the physical fossil record. A key difference, however, is the extent to which information from the abstracted fossil record is disconnected from the physical fossil record. As just one example, genus-level temporal duration is a common variable in many quantitative paleobiological analyses (e.g., Liow, 2007; Powell, 2007; Heim and Peters, 2011). The temporal duration of an individual genus is an amalgamation of the durations of individual species assigned to that genus. In turn, the duration of each species is based on known stratigraphic occurrences of individual specimens hypothesized to belong to that species. Thus, two levels of abstraction—specimens assigned to the species and species assigned to the genus—separate the genus-level duration from the material specimens that support it.

Such issues of the relationships between objects (the physical) and the data (the abstracted) drive much of philosophy in general—"what is the relation between concepts and things?" (Lord, 2006)—and have been discussed by a number of modern philosophers of science (e.g., Daston, 2000, 2004; Hein, 2000; Pinna, 2000). A related central theme in the philosophy of science, especially in recent decades, has been between "realist" views, in which "discoveries" are made about the physical world, and once found are found forever, and "constructionist" or "constructivist" views, which consider data and the interpretations derived from them (or, abstractions) as "inventions," created in the context of culture and biases and requiring occasional returns to the objects subjected to earlier interpretation (e.g., Daston, 2000; Haack, 2007). In the context of paleontology, interpretations based on fossils—such as the assignment of a specimen to a particular species or genus (i.e., the abstracted fossil record) are clearly much more like "constructions" than absolute or indisputable facts (such as the geographic position where a fossil was found or its longest dimension). It is, therefore, perhaps not surprising that the history of paleontology as a science is very much a history of shifting interpretations about the nature of fossils and their implications for the history of the Earth (Rudwick, 1976; Waterston, 1979, p. 10; Cutler, 2003).[2]

If we accept even a partially "constructionist" view of paleontology—in which our physical fossil record (objects) requires constant reinterpretation—then we must recognize specimens themselves as paleontology's only true source of "data,"[3] and the museum collections in which they are housed as the conduit between the two fossil records, allowing testing of past interpretations.[4] Yet paleontological collections are not just repositories for verification of past hypotheses and interpretations. In this essay, we argue that specimen collections and modern tools for digitizing information about them together present an opportunity to expand and transform quantitative paleobiology—which for decades has depended on highly abstracted data derived from an ultimately limited corpus of published literature—into what has come to be called a "big data" science.[5] In such a formulation, greater electronic access to a substantial portion of the millions of fossil specimens currently residing in museum drawers—the "unpublished fossil record" (Teichert et al., 1987)[6]—and the spectrum of potential data associated with them, present

[1]David Sepkoski (2017) has compellingly suggested a slightly different taxonomy: the "original" archive, he says, "is the earth itself," which he labels "archive$_0$." A collection of specimens is "archive$_1$," a pictorial atlas of fossils is "archive$_2$," a "text-only catalog" is "archive$_3$," and a compilation of numerical data—a database—is "archive$_4$." His first two "archives" (0 and 1) are equal to our "physical fossil record," and the next three (2–4) correspond to our "abstracted fossil record."

[2]As Daston noted, paraphrasing Cuvier, "the opposition between secrets of nature laid bare or 'discovered' and 'vast edifices [constructed] on imaginary bases' still haunts our discussions of scientific objects…." (Daston, 2000, p. 5).

[3]We follow Leonelli (2016, p. 77) in defining data as "any product of research activities … that is collected, stored, and disseminated *in order to be used as evidence for knowledge claims*" (emphasis in original). This definition views data as a relational category, which can be assigned to any objects, from physical specimens to the abstracted information about them.

[4]Fossil "collections" consist not just of specimens, but also of the associated human-created structures that make those specimens and their associated information available to researchers. It is these structures—which include physical ordering, safe storage, labeling, and access—that are the "conduit" through which the specimens "flow."

[5]There is no consensus about what "big data" means (Snijders et al., 2012; Mayer-Schönberger and Cukier, 2013). Initially, the idea focused on access to massive amounts of data and the manner in which it was stored and analyzed. The significance of "big data," however, lies not merely in the size of a given database, but in compiling data across previously unassociated databases. For the purposes of this paper, "big data" are of sufficient scale to provide opportunities to find new patterns, ask new questions, and generate new hypotheses that would not have been otherwise feasible or perhaps imaginable. "Big data" will thus likely ultimately be a relative term: what seems "big" today may seem small in the world of tomorrow as our capacity to store, retrieve, and analyze data continues to improve.

[6]The "unpublished" fossil record consists of specimens that have never been mentioned or used in any published scientific work, including systematics, biostratigraphy, or paleoenvironmental reconstructions. Teichert et al. (1987) included in this definition materials mentioned in unpublished theses as prominent examples of fossils that have been studied, but whose abstracted information is nevertheless inaccessible to other researchers.

an opportunity to undertake truly massive quantitative analyses. Such analyses may lead to new patterns, questions, and hypotheses about life's history, but only if we choose to re-embrace the physical fossil record as paleontology's largest and primary source of data. Before considering how paleontology might transform itself into a "big data" science—as well as the challenges that such a transformation presents—we first review how the science of paleontology is philosophically connected to specimens. We then discuss why fossil collections themselves have potentially inexhaustible evidential value, and consider how abstracted data have driven paleobiological research in recent decades.

FOSSILS AND DATA—OBJECTS AND ABSTRACTIONS

Fossils as a Primary Source of Data

Paleontologists study the physical fossil record and, from individual specimens, make empirical observations, which may be recorded as numbers, text, or images that are often called "data." Once these observations are made, the paleontologist usually only works with the derived, abstracted data, rather than the specimen itself. The fossil specimen (the object) is commonly incorporated into a museum collection and stored there long after the paleontologist has finished making her observations. This practice—which differentiates paleontology from sciences such as astronomy, physics, chemistry, and ethology, which do not involve collection of objects—has persisted for centuries, resulting in assembled specimen collections that last well beyond the careers and lifetimes of the paleontologists who first collected and studied them (e.g., Rudwick, 1976; Impey and MacGregor, 1985; Findlen, 1994; Knell, 2000). Paleontologists may also make representational images of the specimens, and these may be gathered together, forming "virtual" (or "paper") collections (e.g., Rudwick, 2000; Davidson, 2008; Hendricks et al., 2015). Both kinds of collections, however, have in common that they contain potentially important information that exists beyond the particular observations made on individual specimens. The process of organizing collections for the simple sake of keeping inventory and retrieving information—for example, assigning identifying numbers to specimens, recording those numbers in a specimen catalog, and, more recently, adding that information to a computerized database—has long been central to museum collections management. Thus, museum collections are the conduit between the physical and abstracted fossil records. Collections are in this way not only archives of previously discovered knowledge, but also reserves of potential future observations—that is, they have great "evidential value."

Evidential Value of Fossil Collections

What makes fossil collections (and the information abstracted from them) into powerful epistemic tools is their ability to act as evidence for a variety of knowledge claims about phenomena across research situations. There are four major reasons why paleontological collections—if they and the information they contain can be made widely available—have nearly inexhaustible evidential value (i.e., effectively unlimited potential for repurposing to serve new scientific goals; sensu Leonelli, 2016).

New Discoveries—New Collections

One of the most important justifications for maintaining large collections in paleontology (and other natural history disciplines) is the potential opportunity for future discoveries using previously collected specimens. There are many examples of older paleontological collections that have yielded new information and sometimes astonishing and unique discoveries (e.g., Allmon, 2005, and references therein; Callomon and Grădinaru, 2005; Boessenkool et al., 2009; Boersma and Pyenson, 2015; Kundrát and Ebbestad, 2015; Lindkvist, 2016). A number of recent studies have also not only discovered something new in existing fossil collections but used them to check previous results, sometimes confirming, supplementing, or disputing published literature data (e.g., Hunter and Donovan, 2005; Davis and Pyenson, 2007; Harnik, 2009; Dean et al., 2010). There are also countless examples of uses of existing fossil collections as sources of data for analyses that were never envisioned when they were first collected (e.g., Parham et al., 2012; Benton et al., 2014; Harnik et al., 2017).

In addition to existing collections, there is still clearly much to be discovered in the world's fossil-bearing rocks, many of which have never been carefully or extensively examined or collected; new collections are therefore clearly justified (e.g., Jackson and Johnson, 2001; Adrain and Westrop, 2003). In some cases, these new collections integrate new kinds of contextual data such as precise geographic coordinates and physicochemical data, and newly collected specimens may shed light on previously made collections.

Verification and/or Record of Observations and Conclusions

The scientific "crown jewels" of many museum fossil collections are "type specimens"—those that formally represent names assigned to new taxa based on the rules of biological nomenclature. This "authority of the specimen" (Waterston, 1979, p. 11) is treated as the central basis for all systematic biology and paleontology. Types are both an "archive of past research" (Knell, 1997) and a basis for future testing of the results of that research. The very concept of the type specimen is premised on the idea of permanently securing the potential for repeatedly comparing specimens with ideas derived from them (Mayr, 1989; Daston, 2004). More broadly, many authors have long argued that the practice of systematic biology itself depends absolutely on the presence of specimen collections, against which hypotheses can be tested (Mayr, 1969, p. 101; Waterston, 1979; Ellis, 2008, p. 173; Sunderland, 2016). "Systematic work that does not rely on museum specimens to verify or falsify the identities of the taxa studied," declared Winston (2007, p. 47) flatly, "is not science."

Because the scholarship of biodiversity includes scrutinising earlier work, evaluating what was written before, and adding new information and insight, it should always be possible to return to those specimens.

They are the primary evidence for the information presented. The ability of researchers to re-examine the primary data and question the conclusions of previous work is a crucial part of what makes this a scientific activity. (Schilthuizen et al., 2015, p. 237)

Replication of Observations

Fundamental to science is the repeatability of results. Yet, unlike most phenomena studied by physicists and chemists, every object studied by paleontologists (and systematic biologists) is historically unique in that it is the product of events that occurred only once in time and space. This means an individual fossil cannot be replicated or easily re-obtained when more observations are needed later. If specimens are not accessible, and similar ones cannot be collected, additional data cannot be gathered. Similarly, it is frequently the case that if fossils are not collected in the field, they cannot be revisited easily later, due to their destruction by natural causes (such as erosion) or human causes (such as development). Many older museum collections contain specimens that simply cannot be collected again because the localities are gone (see examples in Allmon, 2005). Some new collections are strategically made in locations that will or are expected to become inaccessible in the future, even if there is no immediate or pressing scientific reason for doing so (i.e., "salvage paleontology"; e.g., Stone, 2013; Dooley et al., 2015).

Incomplete Nature of Observations

For a variety of reasons, we cannot make all the observations on a fossil at once. First, there may simply be no time to do so in the field. Much paleontological field work is done under conditions that are less than optimal for careful observation. Second, a paleontologist may be unsure of what is "important" to observe. This may be especially the case for unexpected or novel discoveries. Third, many techniques and technologies—from CT scans to isotope sclerochronology—may only be possible in laboratory settings, and many such techniques may become available only long after the fossils are first seen in the field. Fourth, the problems of paleontology change through time: new research questions beget the need for new kinds of observations of previously made collections.

Summary

The evidential value of specimens in collections is large and greatly underdetermined, highlighting the importance of their untapped potential in relation to existing and future research opportunities in paleontology. The nature of the relationship between fossils and the information derived from them has always been complex, but never static. We are always finding new ways to use old collections.

From Fossils to the Paleobiology Database

David Sepkoski (2013) recently argued persuasively that—although paleontology as a science has been "data-driven" almost since its beginnings in the early nineteenth century—there has been a major conceptual shift in the field over the past half-century, particularly "in the way paleontologists have understood the nature of their data and the kinds of knowledge derivable from it: a shift from imagining the record as a visual collection of objects that can be organized in physical or virtual space (the specimen cabinet or the illustrated catalog) to an abstract, randomly accessible collection of data points (a database) existing in information 'cyberspace'" (D. Sepkoski, 2013, p. 439). This recent shift has sharpened the distinction of what the fossil record "is," in both an epistemological sense, as the source of knowledge claims about the history of life, and in an ontological sense, as a collection not only of physical specimens but of "pure data" (D. Sepkoski, 2013, p. 435).[7]

A prime example of this changing view of data in paleontology is "quantitative paleobiology" (sensu Gilinsky and Signor, 1991), which arose in the 1970s and rapidly developed in the 1980s to address some of the largest questions about the history of life, many of which were tied to documentation of Phanerozoic-scale biodiversity patterns (e.g., J. Sepkoski et al., 1981; for historical review, see D. Sepkoski, 2012). The data for these studies—typically compendia of family- (J. Sepkoski, 1982) or genus-level (J. Sepkoski, 2002) taxon durations—were compiled largely from the published systematic literature, rather than directly from specimens. The tradition of using abstracted data from the literature to discover new insights about the history of life continues today. Its most important modern manifestation is the Paleobiology Database (hereafter, PaleoDB[8]), which has become an integral part of the landscape of paleontological research. Begun in the late 1990s, the PaleoDB is the largest single source of paleontological data ever compiled and has offered rich new insights about the history of life (to date, nearly 300 published papers have utilized the PaleoDB as a source of data). Although some of the PaleoDB information is derived from the original fieldwork and taxonomic studies of contributors, the majority of the data have been extracted from published literature (Peters et al., 2014). Various attempts are now under way to greatly expand the reach of the literature that is summarized in the PaleoDB, mainly by advanced information technology (see Peters et al., 2014, for discussion).

As valuable as the PaleoDB has become for discovering major features of life's history (e.g., Uhen et al., 2013), abstracted data compilations from the literature are ultimately finite and

[7]Recently, Tamborini (2015) has drawn attention to another basic philosophical division in paleontological "data." By way of a historical analysis of German paleontology in the late nineteenth century, Tamborini distinguishes between what would later be called the "ideographic" approach—in which paleontology's primary goal is to describe the morphological features of extinct organisms—and the "nomothetic" approach—using paleontological data to generate and test large general theories of evolutionary pattern and causation. We here take as given that the "nomothetic approach" is the appropriate goal of modern paleobiology but agree with Gould (1980), who promulgated these terms in the field, that "[i]deographic factors determine the parameters and then enter as boundary conditions into a nomothetic model" (Gould, 1980, p. 115, *in* Tamborini, 2015, p. 15).

[8]http://paleobiodb.org

dependent wholly on the research interests of the paleontologists who captured primary data through observation of fossil specimens. It is easy to criticize any compilation of information from the abstracted fossil record as incomplete or inadequate (just as it is easy to criticize the physical fossil record as incomplete; Kidwell and Flessa, 1996). Since at least the beginning of the "paleobiological revolution" of the 1970s (D. Sepkoski, 2012), there has been criticism of using large databases to make conclusions about large-scale patterns in the history of life. Skeptics have argued that such analyses make use of data that are not accurate or complete enough, or are otherwise inappropriate, for the purposes to which it has been put (see, for example, Ager, 1988; Boucot, 1990; Prothero and Liter, 2008; Gaskell, 2012, p. 88; Prothero, 2015; and other examples discussed in Raup, 1986, p. 104–105; J. Sepkoski, 1996; Benton, 1999; Miller, 2000, p. 55–56). Advocates of using such data compilations, however, have countered that the point is not the absolute completeness of the information but the shape of the patterns that large quantities of data reveal, even in the face of data error, as long as that error does not contain substantial directional bias; these patterns have been repeatedly tested internally and against supposedly independent data sources (e.g., J. Sepkoski, 1996; Benton, 1999; Adrain and Westrop, 2000; Miller, 2000; Alroy, 2003, 2010; Wagner et al., 2007; Mondal and Harries, 2015). These authors argue that, at least for some major patterns, large-scale patterns have become sufficiently well supported that active hypothesis testing of whether they are broadly correct is effectively finished.

The main argument in support of this conclusion is that additions or corrections to large paleontological databases usually do not significantly alter the large-scale patterns derived from those databases (e.g., J. Sepkoski, 1992, 1996; Benton, 1999; Miller 2000). Yet despite the apparently general consensus that "large-scale" (e.g., global, Phanerozoic) patterns derived from such literature-based databases are robust, questions continue about the utility of these large databases for producing unbiased results about patterns at *smaller* temporal, geographic, taxonomic, and ecological scales. For instance, Vermeij and Leighton (2003), citing examples from mollusks to mammals, argued that the blunt instrument of global diversity analyses unavoidably masks important variation among habitats, regions, and clades. Further, global, Phanerozoic-scale paleobiological analyses often employ higher taxa (e.g., genera) as units of analysis. Such higher taxa—which by their nature are arbitrarily circumscribed—may not always be suitable proxies for species, which in turn are fundamental to understanding many ecological and evolutionary processes (e.g., Hendricks et al., 2014). Regardless of their "reality" in nature, species in practice are defined by specimens (see above), and many of their most interesting ecological and evolutionary qualities are derived from the occurrences and distributions of those specimens in space and time. Even so, published descriptions and other accounts of many species are based upon only a small subset of the specimens that have been collected. New kinds of data that can be linked back to both these published specimens and those that have been collected, but not yet scientifically studied, are needed to better understand linkages—and identify potential disjunctions—between species-level patterns observed at regional scales (ranging from geographic ranges and temporal durations to ecological interactions) with those observed for higher taxa at larger spatiotemporal scales.

PALEONTOLOGY'S "BIG DATA" FUTURE

Which species interactions are most likely to escalate as the climate warms over the next century? How can paleontologists identify species at risk of extinction before they become extinct? When is an ecosystem likely to reach a potentially disastrous tipping point? When will evolution make an invading species spread even faster? Searching for answers to these questions, and many, many more (Seddon et al., 2014), across massively sized data sets (i.e., "big data") has the potential to reveal patterns, trends, and correlations previously unknown, inspiring questions and generating hypotheses (McCulloch, 2013)[9] that could lead to a more "predictive" paleontology.

Our main argument here is that the greatest potential source of new data in paleontology comes not from the corpus of existing literature, but from specimens themselves. Worldwide, paleontological collections hold many tens (probably hundreds) of millions of individual fossils (Allmon, 1997, 2000). Most of these are associated with geographic, stratigraphic, and taxonomic information, as well as data unique to biological individuals, including size, shape, morphological features associated with ontogenetic stage, indicators of taphonomic history, and sometimes records of environmental history (e.g., isotopic data) and past ecological interactions (e.g., predation scars). The data potentially available in these collections, therefore, exist at two levels: the "traditional" specimen-related data of geography, stratigraphy, and taxonomy, and more derivative data, such as chemistry, morphology, taphonomy, or ecology. Furthermore, new kinds of observations and data are invented as new questions and technologies emerge. Each specimen thereby presents an array of potential information for addressing a diversity of paleontological research questions.

Combining such data from millions of specimens in museum collections, and then integrating them with other unique "big data" resources (e.g., high-resolution paleoclimate and paleocean models, paleogeographic reconstructions, genomic data, etc.), has the potential to lead to the most important paleontological discoveries of the twenty-first century. Acquiring and digitizing such large quantities of specimen-based data, mobilizing them

[9]"Big data" research has been criticized for placing too much emphasis on data mining and statistical correlations (Leonelli, 2014). Such concerns reflect deeper debates about the proper role of hypotheses in research that have challenged the scientific community since the seventeenth century (Elliott et al., 2016). Here, we follow Elliott et al. (2016) in thinking of the scientific method as an iteration between both observationally driven and hypothesis-driven modes of investigation.

for dissemination (i.e., making data "travel"[10] so that they can be reused), and linking these data to other ecological and environmental data would be a massive undertaking, but we encourage paleontologists to think imaginatively about what new questions might be answered if—at least in theory—every aspect of every specimen sitting in a museum drawer right now could be captured, digitized, and made freely available for study.

FOUR CHALLENGES

Some steps toward a specimen-based "big data" science in paleontology have already been taken. The Integrated Digitized Biocollections database (iDigBio[11]; see overviews in Page et al., 2015; MacFadden and Guralnick, 2017) has been developed based on data derived directly from specimens themselves, rather than literature sources. Identified museum specimens of known stratigraphic provenance are georeferenced (i.e., their geographic coordinates are determined) and entered into the databases of individual museums before being served to iDigBio; photographs of the fossils themselves are sometimes also captured. These specimen data are then aggregated by iDigBio into a single, online database. New software tools are also being developed to connect the specimen-based neontological and paleontological data aggregated by iDigBio with the literature-based data available in the PaleoDB. For instance, the Enhancing Paleontological and Neontological Data Discovery[12] (ePANDDA) application programming interface (API) enables researchers to access and search these two databases simultaneously, resulting in a more comprehensive data set. Despite their great promise, however, these efforts are still in their infancy, and many challenges remain to realize the full potential of "big data" paleontology. These challenges boil down to the necessity of much larger quantities and greater diversity of high-quality data, as well as greater engagement of taxonomic experts, both of which obviously require much larger investment in collections and human resources.

1. Comprehensiveness

If paleontology is to become a "big data" science, greater attention and energy must be invested in specimens themselves. iDigBio currently contains over 4.6 million individual fossil specimen records, and more than 330,000 of these include additional media such as digital images. Yet this large data set is small in size in comparison to its potential scope, because the vast majority of fossil specimens in collections have not yet been digitized (see Challenge #4). Compounding the problem of small sample size is the reality that specimen data in iDigBio (or any large database) represent a highly selected fraction of what is potentially available in museum collections. All taxonomic groups, geographic regions, and stratigraphic intervals are not adequately represented. Furthermore, many informative aspects of individual specimens are also not included. For instance, the specimen-based data that have been most successfully included into large databases, such as iDigBio, are mainly taxonomic, geographic and/or stratigraphic in nature, using highly standardized vocabularies (Wieczorek et al., 2012). Heterogeneous data that document other morphological, taphonomic, ecological, or geochemical aspects of fossil specimens, however, are typically less standardized in their format, which has made their integration into large databases a challenge (Brewer et al., 2012; see Pauli et al., 2017, for an example with isotopic data). Such bias in large databases exists because what is included in them is not the result of "well-documented scientific choices…. Rather, it is the serendipitous result of social, political, economic, and technical factors, which determine which data get to travel" (Leonelli, 2016, p. 166). Unless paleontology finds a way to improve the comprehensiveness of its "big data," databases that are limited in scope to selected published data or individual museum collections will continue to provide what Leonelli (2016, p. 166) aptly described as a "privileged dissemination platform for a minority of irrationally selected datasets," which encourages an "inherently conservative and implicitly partial platform to discovery."

2. Quality and Reliability of Taxonomic Identifications

Deriving biological meaning from the fossil record—whether viewed as the world's collection of physical specimens or abstracted data that summarize particular aspects of those specimens—ultimately depends on high levels of trust (confidence) among the individual scientists involved. A concrete example is the issue of recognition (and differentiation) of discrete units of evolutionary lineages that many evolutionary biologists would recognize as species (Hendricks et al., 2014; Allmon, 2016). In terms of information content, taxonomic identifications are the metadata that bind specimens both within and across collections, just as a manufacturer's number binds the same product sold by different retailers. Put another way, a taxonomic identification attributed to a fossil specimen is a hypothesis about its phylogenetic position and evolutionary history. Specimen identifications are—along with spatiotemporal context—therefore essential for reconstructing the history of life. Conversely, although identified specimens lacking geographic and temporal context are typically deaccessioned and removed from research collections due to lack of any potential evidential value, those bearing spatiotemporal context—but lacking identifications—are usually stored and protected by museums for decades on end as they await experts to assign them names. Indeed, a digitized specimen record bearing detailed geographic, temporal, ecological, and taphonomic data—but lacking any level of identification—is nearly useless for analyses (even a fossil recognized as "problematic" or *incertae sedis* excludes other possibilities and therefore has some

[10]The dissemination of massive amounts of data is a defining characteristic of the epistemology of "big data" science (Leonelli, 2016). Therefore, the metaphor of data "journey" used by Leonelli (2016, p. 39) is adopted here to designate the "movement of scientific data from their production site to many other sites within or beyond the same field of research."
[11]www.idigbio.org
[12]https://epandda.org/

value). Although this point may seem obvious (just as it is obvious that we could not reconstruct the history of human civilization without the names of key individuals and information about when and where they lived), the problem of taxonomic identifications is not trivial. One of the greatest challenges for paleontology in the age of "big data" is the same challenge that has always vexed museum collections: attaining correct identifications for individual fossil specimens.

A large proportion of the fossil specimens currently residing in museum collections are not identified to the species level, and many that have been assigned names are misidentified. The effort required to collect specimens in the field is often much less than the effort needed to identify those specimens, especially at lower taxonomic levels. Generally, however, names are assigned to specimens—often long after they were first collected—by comparing them with images in printed literature. In this respect, specimen-based paleontology remains mired in the technology of the nineteenth century. Many fossil species have been figured only a small number of times, frequently in images of poor quality by modern standards, in publications that are out of print or otherwise difficult to attain (though digital scans made available through initiatives such as the Biodiversity Heritage Library[13] are now making some of these works more widely accessible). Comparisons between specimens and images in the printed literature available on hand may result in identifications that are correct at higher taxonomic levels but erroneous at lower levels. Paleontology needs new, free digital resources to help both professional and avocational paleontologists identify fossils; the new Digital Atlas of Ancient Life[14] project presents one example for how this might be accomplished (Hendricks et al., 2015).

Misidentifications of fossils and nomenclatural conflicts (e.g., synonymous species names or differences of opinion concerning generic assignments) often result in significant quantities of "messy" taxonomic data in large databases compiled from specimen-based data. Paleontology desperately needs resources such as the World Registry of Marine Species (WoRMS[15]), which seeks "to provide an authoritative and comprehensive list of names of marine organisms, including information on synonymy."[16] Although the PaleoDB allows for entry and accumulation of taxonomic opinion,[17] it was never intended to serve as a comprehensive nomenclatural platform for paleontology; rather, it offers a helpful but incomplete reflection of the taxonomic opinions contained within the literature. The further development of a single common listing could provide nomenclatural consistency across museum collections and tie into larger specimen-based databases, allowing universal updates to taxonomic names where necessary. With expanded active involvement of taxonomic experts, the PaleoDB could potentially evolve to meet this need, but development of a dedicated taxonomic database for paleontology is also worthy of consideration.[18]

Mayer-Schönberger and Cukier (2013) rejected the epistemological significance of such "messiness" of large databases, claiming that "big data" is self-correcting because of its diversity and variability. In a world of small data, reducing bias and errors and ensuring high quality of data are essential. Large data sets, however, make it likely that errors are automatically reduced by a process philosophers call "triangulation"—the tendency of reliable data to cluster together (Wylie, 2002; Leonelli, 2014). We agree, however, with Leonelli (2014) that if data sources share the same bias, it is also possible that bias will be amplified, rather than reduced, especially when a database is far from being comprehensive, which, as discussed above, is presently the case in paleontology—and will be for the foreseeable future.

3. Lack of Engagement by Researchers in Data Curation Activities

Realizing the full evidential value of museum collections in "big data" science requires the support and cooperation of the broader paleontological research community. Such support, however, is currently not encouraged because the scientific credit system typically does not place a high value on curatorial efforts by researchers to improve museum collections or the contribution by researchers to museum specimen collections and data outside of the publications stemming from research. For example, a 2002 National Research Council (NRC) report on the status and future of geoscience collections specifically recommended development of systems for rewarding care and use of collections, but little progress has been made to date (NRC, 2002). Many "curators" in large museums today perform little actual curation—now usually called "collections management"—activity; they are instead researchers who may or may not use, or even know much about, the collections under their nominal supervision. Incentivizing researchers to engage more with collections could help address this challenge.

4. Availability of Funding to "Digitize Everything"

Huge amounts of labor are needed to make data "travel" (i.e., to enable it to be reused and gain new evidential value) (Leonelli, 2016), starting with gathering the basic data in digital form. Funding for digital database creation and curation, however, remains scarce relative to the manifest need. Some crude calculations can provide a sense of the scope of this challenge. It is likely that the 20 or so largest museum collections of invertebrate fossils in the United States contain collectively *at least* 75 million specimens (Allmon and White, 2000). If we assume that ~5% of these specimens are currently represented by digital records, that ~50% are

[13]https://www.biodiversitylibrary.org/
[14]http://www.digitalatlasofancientlife.org
[15]http://www.marinespecies.org/
[16]http://www.marinespecies.org/about.php
[17]paleobiodb.org/data1.2/opinions_doc.html

[18]We recognize that there is a certain irony in the fact that a discipline sometimes pejoratively characterized as "stamp collecting" has no comprehensive listing of fossil species.

already identified, and that it costs approximately US$1 to digitize each specimen, we can then estimate that it would cost perhaps US$35 million to digitize all of the currently *identified* fossil specimens, and more than US$75 million to digitize all fossil specimens in U.S. collections.[19] By comparison, the U.S. National Science Foundation's 2017 budget for digitization of its Advancing Digitization of Biological Collections (ADBC) program for *all* natural history collections was US$10 million.[20]

These numbers make it obvious that digitizing all or even most fossils in museum collections is highly unlikely to occur in the near future. If this is true, then other questions arise, including how much digitization is sufficient (and for what purpose) and which specimens and associated data will be prioritized over others? These are vital questions for the future of museums, collections, and paleontological research.

CONCLUSIONS

Specimen collections are, have been, and shall forever be the epistemological engine of paleontology, offering effectively inexhaustible evidential value. Indeed, the physical fossil record is paleontology's only path toward a truly "big data" future, which may well present insights that we cannot currently predict and lead us to new questions that we cannot currently imagine asking. While adding more specimens with quality taxonomic and spatiotemporal data to databases such as iDigBio should always be a central digitization priority for museum collections, our discipline's "big data" future also resides in other aspects of the abstracted fossil record that generally go unrecorded in existing online databases (examples include morphological, ecological, taphonomic, and geochemical data). By its nature, the published literature is not designed to and cannot make available such "dark data." Excluding this kind of information risks a conceptual conservatism in the questions we can ask about the fossil record.

Moving our discipline in this direction will require considerable effort and resources. But, there are places where we can begin this transformation now. For example, many collections remain inadequately studied and represented in the large distributed databases that are increasingly important research tools in paleontology. Collections might therefore profitably focus their data curation efforts on areas that could be identified as less well represented in large databases, such as particular geologic ages, taxonomic groups, or geographic regions. Making taxonomy more reliable and consistent in order to increase the fitness of data is another major challenge for collections. Collections care might therefore be focused on taxa that are in particular need of clarification. For collections to maximize their contribution to databases in an age of "big data" science, they must also somehow be mobilized to release more of the information that they contain. This means more than just "putting them online" and should also include getting more researchers into the collections to study them, identify material, and sort what is important from what is not.

The "big data" future of paleontology is in many ways the next logical phase of the exploratory role that natural history museums and their collections have always played. "Data curators" (e.g., Freitas and Curry, 2016; Leonelli, 2016) are now becoming just as important as specimen curators for assembling big data and making it possible to integrate and analyze them. Such expanded roles for data and the people who curate them will not replace museum specimen collections. Instead, they will make these collections even more valuable and widely accessible.

Physical specimens have always formed the first pillar of the fossil record. We suggest that the untold quantities of data that are potentially abstractable from the physical fossil record could provide a twenty-first-century upgrade to the second pillar, which modern paleontology remains balanced near, albeit staidly.

ACKNOWLEDGMENTS

We thank Gary Rosenberg and Renee Clary for inviting us to participate in the symposium and volume; Amanda Schmitt for assistance with references; Erica Clites, Kirk Johnson, Bruce Lieberman, and Austin Hendy for discussion; and Mike Benton and an anonymous reviewer for comments on previous drafts.

REFERENCES CITED

Adrain, J.M., and Westrop, S.R., 2000, An empirical assessment of taxic paleobiology: Science, v. 289, p. 110–112, https://doi.org/10.1126/science.289.5476.110.

Adrain, J.M., and Westrop, S.R., 2003, Paleobiodiversity: We need new data: Paleobiology, v. 29, p. 22–25, https://doi.org/10.1666/0094-8373(2003)029<0022:PWNND>2.0.CO;2.

Ager, D.V., 1988, Extinctions and survivals in the Brachiopoda and the dangers of data bases, *in* Larwood, G.P., ed., Extinction and Survival in the Fossil Record: Oxford, UK, Oxford University Press, p. 89–97.

Allmon, W.D., 1997, Collections in paleontology, *in* Lane, H.R., Lipps, J., Steininger, F.F., and Ziegler, W., eds., Paleontology in the 21st Century Workshop: Kleine Senckenbergreihe, no. 25, p. 155–159.

Allmon, W.D., 2000, Collections, *in* Lane, H.R., Lipps, J., Steininger, F.F., Kaesler, R.L. Ziegler, W., and Lipps, J., eds., Fossils and the Future, Paleontology in the 21st Century: Frankfurt, Senckenberg-Buch, no. 74, p. 203–214.

Allmon, W.D., 2005, The importance of museum collections in paleobiology: Paleobiology, v. 31, p. 1–5.

Allmon, W.D., 2016, Studying species in the fossil record: A review and recommendations for a more unified approach, *in* Allmon, W.D., and Yacobucci, M.M., eds., Species and Speciation in the Fossil Record: Chicago, University of Chicago Press, p. 59–120, https://doi.org/10.7208/chicago/9780226377582.001.0001.

[19]These cost estimates are based on our participation in three paleo-themed "Thematic Collections Network" grants (PaleoNiches, Eastern Pacific Invertebrate Communities of the Cenozoic [EPICC], and Cretaceous World projects) funded by the U.S. National Science Foundation's Advancing Digitization of Biodiversity Collections (ADBC) program. Our experience is that digitization costs are extremely variable among institutions and are surprisingly poorly understood by both institutions and funding agencies. This situation appears to be the result of a combination of differences in work flow, staffing, institutional infrastructure, and details of methods and technology. The "$1 per specimen" is the estimated average cost of digitally capturing "basic label data" and georeferencing for one specimen or set of specimens. Importantly, this estimate does *not* include any imaging of specimens, which our experience suggests costs significantly more.

[20]https://www.nsf.gov/pubs/2015/nsf15576/nsf15576.pdf

Allmon, W.D., and White, R.D., 2000, Introduction, *in* White, R.D., and Allmon, W.D., eds., Guidelines for the Management and Curation of Invertebrate Fossil Collections: Paleontological Society Special Publications, v. 10, p. 1–4.

Alroy, J., 2003, Global databases will yield reliable measures of global biodiversity: Paleobiology, v. 29, p. 26–29, https://doi.org/10.1666/0094-8373(2003)029<0026:GDWYRM>2.0.CO;2.

Alroy, J., 2010, Fair sampling of taxonomic richness and unbiased estimation of origination and extinction rates, *in* Alroy, J., and Hunt, G., eds., Quantitative Methods in Paleobiology: The Paleontological Society Papers, v. 16, p. 55–80.

Benton, M., 1999, The history of life: Large databases in palaeontology, *in* Harper, D.A., ed., Numerical Palaeobiology: Computer-based Modelling and Analysis of Fossils and Their Distributions: New York, John Wiley, p. 249–283.

Benton, M.J., Forth, J., and Langer, M.C., 2014, Models for the rise of the dinosaurs: Current Biology, v. 24, p. R87–R95, https://doi.org/10.1016/j.cub.2013.11.063.

Boersma, A.T., and Pyenson, N.D., 2015, *Albicetus oxymycterus,* a new generic name and redescription of a basal physeteroid (Mammalia, Cetacea) from the Miocene of California, and the evolution of body size in sperm whales: PLoS One, v. 10, p. e0135551, https://doi.org/10.1371/journal.pone.0135551.

Boessenkool, S., Austin, J.J., Worthy, T.H., Scofield, P., Cooper, A., Seddon, P.J., and Waters, J.M., 2009, Relict or colonizer?: Extinction and range expansion of penguins in southern New Zealand: Proceedings of the Royal Society B: Biological Sciences, v. 276, p. 815–821, https://doi.org/10.1098/rspb.2008.1246.

Boucot, A.J., 1990, Phanerozoic extinctions: How similar are they to each other?, *in* Kauffman, E.G., and Walliser, O.H., eds., Extinction Events in Earth History: Berlin, Heidelberg, Lecture Notes in Earth Sciences, v. 30, Springer, p. 5–30, https://doi.org/10.1007/BFb0011131.

Brewer, S., Jackson, S.T., and Williams, J.W., 2012, Paleoecoinformatics: Applying geohistorical data to ecological questions: Trends in Ecology & Evolution, v. 27, p. 104–112, https://doi.org/10.1016/j.tree.2011.09.009.

Callomon, J.H., and Grădinaru, E., 2005, From the thesaurus of the museum collections. I. Liassic ammonites from Munteana (Svinita Zone, Southern Carpathians, Romania): Acta Palaeontologia Romaniae, v. 5, p. 49–65.

Cutler, A., 2003, The Seashell on the Mountaintop: New York, Dutton, 240 p.

Daston, L., 2000, Introduction: The coming into being of scientific objects, *in* Daston, L., ed., Biographies of Scientific Objects: Chicago, University of Chicago Press, p. 1–14.

Daston, L., 2004, Type specimens and scientific memory: Critical Inquiry, v. 31, p. 153–182, https://doi.org/10.1086/427306.

Davidson, J.P., 2008, A History of Paleontology Illustration: Bloomington, Indiana University Press, 217 p.

Davis, E.B., and Pyenson, N.D., 2007, Diversity biases in terrestrial mammalian assemblages and quantifying the differences between museum collections and published accounts: A case study from the Miocene of Nevada: Palaeogeography, Palaeoclimatology, Palaeoecology, v. 250, p. 139–149, https://doi.org/10.1016/j.palaeo.2007.03.006.

Dean, M.T., Owen, A.W., Bowdler-Hicks, A., and Akhurst, M.C., 2010, Discriminating faunal assemblages and their palaeoecology based on museum collections: The Carboniferous Hurlet and Index limestones of western Scotland: Scottish Journal of Geology, v. 46, p. 45–57, https://doi.org/10.1144/0036-9276/01-399.

Dooley, A.C., Jr., Heckert, A.B., Fraser, N.C., Byrd, C.J., Vodden, R., and Beard, J., 2015, Salvage paleontology at the Virginia Solite Quarry (Upper Triassic)—Rescuing a lagerstätte: Geological Society of America Abstracts with Programs, v. 47, no. 2, p. 24.

Elliott, K.C., Cheruvelil, K.S., Montgomery, G.M., and Soranno, P.A., 2016, Conceptions of good science in our data-rich world: Bioscience, v. 66, p. 880–889, https://doi.org/10.1093/biosci/biw115.

Ellis, R., 2008, Rethinking the value of biological specimens: Laboratories, museums and the Barcoding of Life Initiative: Museum & Society, v. 6, no. 2, p. 172–191.

Findlen, P., 1994, Possessing Nature: Museums, Collecting, and Scientific Culture in Early Modern Italy: Berkeley, University of California Press, 449 p.

Freitas, A., and Curry, E., 2016, Big data curation, *in* Cavanillas, J.M., et al., eds., New Horizons for a Data-Driven Economy: New York, Springer, p. 87–118, https://doi.org/10.1007/978-3-319-21569-3_6

Gaskell, I., 2012, Museums and philosophy—Of art, and many other things, Part II: Philosophy Compass, v. 7, no. 2, p. 85–102, https://doi.org/10.1111/j.1747-9991.2011.00469.x.

Gilinsky, N.L., and Signor, P.W., eds., 1991, Analytical Paleobiology: The Paleontological Society, Short Courses in Paleontology, no. 4, 216 p.

Gould, S.J., 1980, The promise of paleobiology as a nomothetic, evolutionary discipline: Paleobiology, v. 6, no. 1, p. 96–118.

Haack, S., 2007, Defending Science—Within Reason. Between Scientism and Cynicism: Amherst, New York, Prometheus Books, 411 p.

Harnik, P.G., 2009, Unveiling rare diversity by integrating museum, literature, and field data: Paleobiology, v. 35, no. 2, p. 190–208, https://doi.org/10.1666/07062.1.

Harnik, P.G., Serb, J.M., and Sherratt, E., 2017, Morphological evolution following the closure of the Central American Seaway: Geological Society of America Abstracts with Programs, v. 49, no. 6, https://doi.org/10.1130/abs/2017AM-301854.

Heim, N.A., and Peters, S.E., 2011, Regional environmental breadth predicts geographic range and longevity in fossil marine genera: PLoS One, v. 6, no. 5, p. e18946.

Hein, H.S., 2000, The Museum in Transition. A Philosophical Perspective: Washington, D.C., Smithsonian Institution Press, 203 p.

Hendricks, J.R., Saupe, E.E., Myers, C.E., Hermsen, E.J., and Allmon, W.D., 2014, The generification of the fossil record: Paleobiology, v. 40, no. 4, p. 511–528.

Hendricks, J.R., Stigall, A.L., and Lieberman, B.S., 2015, The Digital Atlas of Ancient Life: Delivering information on paleontology and biogeography via the web: Palaeontologia Electronica, v. 18, no. 2, p. 1–9.

Hunter, A., and Donovan, S., 2005, Field sampling bias, museum collections and completeness of the fossil record: Lethaia, v. 38, no. 4, p. 305–314, https://doi.org/10.1080/00241160500289559.

Impey, O., and MacGregor, A., 1985, The Origins of Museums: The Cabinet of Curiosities in Sixteenth- and Seventeenth-Century Europe: Oxford, UK, Oxford University Press, 335 p.

Jackson, J.B., and Johnson, K.G., 2001, Measuring past biodiversity: Science, v. 293, p. 2401–2404, https://doi.org/10.1126/science.1063789.

Kidwell, S.M., and Flessa, K.W., 1996, The quality of the fossil record: Populations, species, and communities: Annual Review of Earth and Planetary Sciences, v. 24, no. 1, p. 433–464, https://doi.org/10.1146/annurev.earth.24.1.433.

Knell, S.J., 1997, What's important?, *in* Pettitt, C.W., and Nudds, J.R., eds., The Value and Valuation of Natural Science Collections: London, the Geological Society, p. 11–16.

Knell, S.J., 2000, The Culture of English Geology 1815–1851: A Science Revealed through Its Collecting: Aldershot, UK, Ashgate, 377 p.

Kundrát, M.L.W., and Ebbestad, J.O.R., 2015, New tooth of Peking Man recognized in laboratory in Uppsala University: Acta Anthropologica Sinica, v. 34, no. 1, p. 1–14.

Leonelli, S., 2014, What difference does quantity make?: On the epistemology of Big Data in biology: Big Data & Society, April–June 2014, p. 1–11.

Leonelli, S., 2016, Data-Centric Biology: A Philosophical Study: Chicago, University of Chicago Press, 275 p.

Lindkvist, M., 2016, The importance of curation: A case-study of the subfossil lemur collection in the Museum of Evolution [Ph.D. thesis]: Uppsala, Sweden, Uppsala University, 84 p.

Liow, L.H., 2007, Does versatility as measured by geographic range, bathymetric range, and morphological variability contribute to taxon longevity?: Global Ecology and Biogeography, v. 16, p. 117–128, https://doi.org/10.1111/j.1466-8238.2006.00269.x.

Lord, B., 2006, Philosophy and the museum: An introduction to the special issue: Museum Management and Curatorship, v. 21, p. 79–87, https://doi.org/10.1080/09647770600102102.

MacFadden, B.J., and Guralnick, R.P., 2017, Horses in the cloud: Big data exploration and mining of fossil and extant *Equus* (Mammalia: Equidae): Paleobiology, v. 43, no. 1, p. 1–14, https://doi.org/10.1017/pab.2016.42.

Mayer-Schönberger, V., and Cukier, K., 2013, Big Data: A Revolution That Will Transform How We Live, Work and Think: New York, Houghton Mifflin Harcourt, 242 p.

Mayr, E., 1969, Principles of Systematic Zoology: New York, McGraw-Hill, 434 p.

Mayr, E., 1989, Attaching names to objects, *in* Ruse, M., ed., What the Philosophy of Biology Is: Essays for David Hull: Dordrecht, Netherlands, Kluwer Academic, p. 235–243, https://doi.org/10.1007/978-94-009-1169-7_12.

McCulloch, E.S., 2013, Harnessing the power of big data in biological research: Bioscience, v. 63, no. 9, p. 715–716, https://doi.org/10.1093/bioscience/63.9.715.

Miller, A.I., 2000, Conversations about Phanerozoic global diversity, *in* Erwin, D.H., and Wing, S.L., eds., Deep Time. Paleobiology's Perspective: Paleobiology 26 (Suppl. to No. 4), p. 53–73.

Mondal, S., and Harries, P.J., 2015, The effect of taxonomic corrections on Phanerozoic generic richness trends in marine bivalves with a discussion of the clade's overall history: Paleobiology, v. 42, no. 1, p. 157–171, https://doi.org/10.1017/pab.2015.35.

National Research Council (NRC) (Committee on the Preservation of Geoscience Data and Collections), 2002, Geoscience Data and Collections: National Resources in Peril: Washington, D.C., The National Academies Press, 107 p.

Page, L.M., MacFadden, B.J., Fortes, J.A., Soltis, P.S., and Riccardi, G., 2015, Digitization of biodiversity collections reveals biggest data on biodiversity: Bioscience, v. 65, no. 9, p. 841–842, https://doi.org/10.1093/biosci/biv104.

Parham, J.F., Donoghue, P.C.J., Bell, C.J., Calway, T.D., Head, J.J., Holroyd, P.A., Inoue, J.G., Irmis, R.B., Joyce, W.G., Ksepka, D.T., Patané, J.S.L., Smith, N.D., Tarver, J.E., van Tuinen, M., Yang, Z., Angielczyk, K.D., Greenwood, J.M., Hipsley, C.A., Jacobs, L., Makovicky, P.J., Müller, J., Smith, K.T., Theodor, J.M., Warnock, R.C.M., and Benton, M.J., 2012, Best practices for justifying fossil calibrations: Systematic Biology, v. 61, no. 2, p. 346–359, https://doi.org/10.1093/sysbio/syr107.

Pauli, J.N., Newsome, S.D., Cook, J.A., Harrod, C., Steffan, S.A., Backer, C.J.O., Ben-David, M., Bloom, D., Bowen, G.J., Cerling, T.E., Cicero, C., Cook, C., Dohm, M., Dharampal, P.S., Graves, G., Gropp, R., Hobson, K.A., Jordan, C., MacFadden, B., Pilaar Birch, S., Poelen, J., Ratnasingham, S., Russell, L., Stricker, C.A., Uhen, M.D., Yarnes, C.T., and Hayden, B., 2017, Why we need a centralized repository for isotopic data: Proceedings of the National Academy of Sciences of the United States of America, v. 114, no. 12, p. 2997–3001, https://doi.org/10.1073/pnas.1701742114.

Peters, S.E., Zhang, C., Livny, M., and Ré, C., 2014, A machine reading system for assembling synthetic paleontological databases: PLoS One, v. 9, p. e113523, https://doi.org/10.1371/journal.pone.0113523.

Pinna, G., 2000, A philosophy for natural history museums, *in* Ghiselin, M.T., and Leviton, A.E., eds., Cultures and Institutions of Natural History: San Francisco, California Academy of Sciences, p. 333–337.

Powell, M.G., 2007, Geographic range and genus longevity of late Paleozoic brachiopods: Paleobiology, v. 33, no. 4, p. 530–546, https://doi.org/10.1666/07011.1.

Prothero, D.R., 2015, Garbage in, garbage out: The effect of immature taxonomy on database compilations of North American fossil mammals, *in* Sullivan, R.M., and Lucas, S.G., eds., Fossil Record 4: New Mexico Museum of Natural History and Science Bulletin, v. 68, p. 257–264.

Prothero, D.R., and Liter, M.R., 2008, Systematics of the dromomerycines and aletomerycines (Artiodactyla: Palaeomerycidae) from the Miocene and Pliocene of North America, *in* Lucas, S.G., Morgan, G.S., Spielmann, J.A. and Prothero, D.R., eds., Neogene Mammals: New Mexico Museum of Natural History and Science Bulletin, v. 44, p. 273–298.

Raup, D.M., 1986, The Nemesis Affair: A Story of the Death of Dinosaurs and the Ways of Science: New York, W.W. Norton, 220 p.

Rudwick, M.J.S., 1976, The Meaning of Fossils: Episodes in the History of Palaeontology: 2nd edition: Chicago, University of Chicago Press, 287 p.

Rudwick, M.J.S., 2000, Georges Cuvier's paper museum of fossil bones: Archives of Natural History, v. 27, no. 1, p. 51–68, https://doi.org/10.3366/anh.2000.27.1.51.

Schilthuizen, M., Vairappan, C.S., Slade, E.M., Mann, D.J., and Miller, J.A., 2015, Specimens as primary data: Museums and 'open science': Trends in Ecology & Evolution, v. 30, no. 5, p. 237–238, https://doi.org/10.1016/j.tree.2015.03.002.

Seddon, A.W.R., Mackay, A.W., Baker, A.G., Birks, H.J.B., Breman, E., Buck, C.E., Ellis, E.C., Froyd, C.E., Gill, J.L., Gillson, L., Johnson, E.A., Jones, V.J., Juggins, S., Macias-Fauria, M., Mills, K., Morris, J.L., Nogués-Bravo, D., Punyasena, S.W., Roland, T.P., Tanentzap, A.J., Willis, K.J., Aberhan, M., van Asperen, E.N., Austin, W.E.N., Battarbee, R.W., Bhagwat, S., Belanger, C.L., Bennett, K.D., Birks, H.H., Bronk Ramsey, C., Brooks, S.J., de Bruyn, M., Butler, P.G., Chambers, F.M., Clarke, S.J., Davies, A.L., Dearing, J.A., Ezard, T.H.G., Feurdean, A., Flower, R.J., Gell, P., Hausmann, S., Hogan, E.J., Hopkins, M.J., Jeffers, E.S., Korhola, A.A., Marchant, R., Kiefer, T., Lamentowicz, M., Larocque-Tobler, I., López-Merino, L., Liow, L.H., McGowan, S., Miller, J.H., Montoya, E., Morton, O., Nogué, S., Onoufriou, C., Boush, L.P., Rodriguez-Sanchez, F., Rose, N.L., Sayer, C.D., Shaw, H.E., Payne, R., Simpson, G., Sohar, K., Whitehouse, N.J., Williams, J.W., and Witkowski, A., 2014, Looking forward through the past: Identification of 50 priority research questions in palaeoecology: Journal of Ecology, v. 102, p. 256–267, https://doi.org/10.1111/1365-2745.12195.

Sepkoski, D., 2012, Rereading the Fossil Record: The Growth of Paleobiology as an Evolutionary Discipline: Chicago, University of Chicago Press, 440 p., https://doi.org/10.7208/chicago/9780226748580.001.0001.

Sepkoski, D., 2013, Towards "a natural history of data": Evolving practices and epistemologies of data in paleontology 1800–2000: Journal of the History of Biology, v. 46, p. 401–444, https://doi.org/10.1007/s10739-012-9336-6.

Sepkoski, D., 2017, The Earth as archive: Contingency, narrative, and the history of Life, *in* Daston, L., ed., Science in the Archives: Pasts, Presents, Futures: Chicago, University of Chicago Press, p. 53–83.

Sepkoski, J.J., Jr., 1982, A compendium of marine animal families: Milwaukee Public Museum Contributions to Biology and Geology, v. 51, no. 1, p. 1–125.

Sepkoski, J.J., Jr., 1992, Phylogenetic and ecologic patterns in the Phanerozoic history of marine biodiversity, *in* Eldredge, N., ed., Systematics, Ecology, and the Biodiversity Crisis: New York, Columbia University Press, p. 77–100.

Sepkoski, J.J., Jr., 1996, Patterns of Phanerozoic extinction: A perspective from global data bases, *in* Walliser, O., ed., Global Events and Event Stratigraphy in the Phanerozoic: Berlin, Springer, p. 35–51, https://doi.org/10.1007/978-3-642-79634-0_4.

Sepkoski, J.J., Jr., 2002, A compendium of fossil marine animal genera: Bulletins of American Paleontology, no. 363, 560 p.

Sepkoski, J.J., Jr., Bambach, R.K., Raup, D.M., and Valentine, J.W., 1981, Phanerozoic marine diversity and the fossil record: Nature, v. 293, p. 435–437, https://doi.org/10.1038/293435a0.

Snijders, C., Matzat, U., and Reips, U.D., 2012, Big data: Big gaps of knowledge in the field of internet science: International Journal of Internet Science, v. 7, p. 1–5.

Stone, R., 2013, Salvage paleontology on the seaway: Science, v. 341, no. 6143, p. 232, https://doi.org/10.1126/science.341.6143.232.

Sunderland, M.E., 2016, Specimens and collections, *in* Lightman, B., ed., A Companion to the History of Science: West Sussex, UK, John Wiley & Sons Ltd., p. 488–499.

Tamborini, M., 2015, The constitution of paleobiological data [Ph.D. dissertation]: Germany, University of Heidelberg, 264 p.

Teichert, C., Sweet, W.C., and Boucot, A.J., 1987, The unpublished fossil record: Implications: Senckenbergiana Lethaea, v. 68, no. 1–4, p. 5–19.

Uhen, M.D., Barnosky, A.D., Bills, B., Blois, J., Carrano, M.T., Carrasco, M.A., Erickson, G.M., Eronen, J.T., Fortelius, M., Graham, R.W., and Grimm, E.C., 2013, From card catalogs to computers: Databases in vertebrate paleontology: Journal of Vertebrate Paleontology, v. 33, no. 1, p. 13–28.

Vermeij, G.J., and Leighton, L.R., 2003, Does global diversity mean anything?: Paleobiology, v. 29, p. 3–7, https://doi.org/10.1666/0094-8373(2003)029<0003:DGDMA>2.0.CO;2.

Wagner, P.J.M., Aberhan, M., Hendy, A., and Kiessling, W., 2007, The effects of taxonomic standardization on sampling-standardized estimates of historical diversity: Proceedings of the Royal Society B: Biological Sciences, v. 274, no. 1608, p. 439–444, https://doi.org/10.1098/rspb.2006.3742.

Waterston, C.D., 1979, The unique role of the curator in palaeontology, *in* Bassett, M.G., ed., Curation of Palaeontological Collections: Special Papers in Palaeontology, no. 22, p. 7–15.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Doring, M., Giovanni, R., Robertson, T., and Vieglais, D., 2012, Darwin Core: An evolving community-developed biodiversity data standard: PLoS One, v. 7, no. 1, p. e29715, https://doi.org/10.1371/journal.pone.0029715.

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

Wylie, A., 2002, Thinking from Things: Essays in the Philosophy of Archeology: Berkeley, University of California Press, 339 p.

Manuscript Accepted by the Society 30 November 2017

Printed in the USA