Horses in the Cloud: big data exploration and mining of fossil and extant Equus (Mammalia: Equidae)

Authors: MacFadden, Bruce J., and Guralnick, Robert P.

Source: Paleobiology, 43(1) : 1-14

Published By: The Paleontological Society

URL: https://doi.org/10.1017/pab.2016.42

Paleobiology, 43(1), 2017, pp. 1–14
DOI: 10.1017/pab.2016.42

Paleobiology Letters
RAPID COMMUNICATION

Horses in the Cloud: big data exploration and mining of fossil and extant *Equus* (Mammalia: Equidae)

Bruce J. MacFadden and Robert P. Guralnick

Abstract.—Extant species of the genus *Equus* (e.g., horses, asses, and zebras) have a widespread distribution today on all continents except Antarctica. Extinct species of *Equus* represented by fossils were likewise widely distributed in the Pliocene and even more so during the Pleistocene. In order to understand the efficacy of "big data" for (paleo)biogeographic analyses, location records (latitude, longitude) and fossil occurrences for the genus *Equus* were mined and further explored from six databases, including iDigBio, Paleobiology Database, VertNet, BISON, Neotoma, and GBIF. These were chosen from a priori knowledge of where relevant data might be aggregated. We also realized that these databases have different objectives and data sources and therefore would provide a useful comparative study of the widespread taxon *Equus* in space and time.

The mining of *Equus* data from these six sources yielded a combined total of 123.8 K location records, including 116.2 K fossil specimens. These include individual points that are unique, that is, only occurring in one of these databases, and those that are duplicated in multiple databases. Of the six databases, three (iDigBio, Paleobiology Database, and GBIF) were judged to be the most useful in the *Equus* use case. Most of the databases are biased toward North American records, thus limiting the reconstruction of the actual distribution of the genus *Equus* in space and time outside of this continent. Although *Equus* has a large number of digitally accessible records, fundamentally interesting questions pertaining to evolutionary dynamics and extinction geography are still a challenge for these kinds of biodiversity databases due primarily to the lack of sufficiently dense and precise temporal data.

Bruce J. MacFadden and Robert P. Guralnick. Florida Museum of Natural History, University of Florida, Gainesville, Florida 32611, U.S.A. E-mail: bmacfadd@flmnh.ufl.edu

Accepted: 13 September 2016
Published online: 21 October 2016
Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.qc2fm

Introduction

Over the past decade the development of two separate innovations, that is, assembling data in common formats (Constable et al. 2010) and aggregating these data to Web-enabled platforms (Wolstencroft et al. 2013), have revolutionized the kinds and quantity of data that can be accessed and the questions that can be asked about biodiversity and related subjects. For any given taxon, it can be unclear which of these online resources should be accessed for specific research questions. In addition, this field is growing so quickly that some of the data platforms and services that are relevant today may soon become subsumed or obsolete tomorrow. Standardized data sets, while an advantage for broad sharing across platforms, can also lead to end-users accessing different duplicated records, thus potentially complicating analyses. In this paper we present the results of data discovery and mining for the fossil and extant equid (e.g., horses, asses, and zebras) *Equus*. This paper is intended as a metaresearch study, that is, the scientific examination of how research is designed, carried out, communicated, and evaluated (Kousta et al. 2016), followed by an essay on the status and future of big biodiversity databases, particularly those that seek to integrate fossil and extant data.

Thus, for this study we focus on a more complete understanding of the biogeographic distribution of this genus, including: (1) location for all records, which is represented by latitude, longitude, or the equivalent decimalLatitude and decimalLongitude in the DwC (a list of abbreviations and acronyms used in this paper is presented in the Appendix) data standards (Wieczorek et al. 2012); and (2) the DwC designation of basisOfRecord as FossilSpecimen. Location data also happen to be some of the most frequently recorded

0094-8373/17

attributes of individual records, whether these represent vouchered specimens in museum collections or observations made by amateur or expert observers ("citizen scientists"). In the same way, fossil data are fundamentally important for the *Equus* use case, although despite the large number of records retrieved during this study, the temporal data across platforms are not sufficiently consistent or dense to answer meaningful queries such as those related to evolutionary dynamics and extinction geography.

We assert that *Equus* is a good example, or use case, of the kinds of questions and concerns that potentially arise when accessing big biodiversity databases, particularly ones that include both fossil and extant data. It also exemplifies the complexity of data discovery and mining and finding solutions to quality control and integration of results from multiple databases. In addition, we have chosen *Equus* for several reasons, including the following:

1. The genus *Equus* L. 1758 is a taxonomically stable name and as such is not prone to some of the complexities of other taxa for which the nomen has changed since its original description. Although there is some debate about subgenera, we assert that the vast majority of researchers using the nomen *Equus* have little argument about its identity. (The validity and allocation of the 8 to 10 generally recognized extant species within the genus *Equus* have undergone considerable changes, but this fact is not of concern for the goals of this study.)
2. The genus is extant and also has a widespread fossil record, with the latter including many named species and considerable discussion with regard to the validity of individual specific nomina. As such, we predicted a priori that we would need to access multiple databases to fully understand the location of *Equus* in space and time.
3. When dealing with the paleobiogeographic distribution of fossil taxa, past continental configurations typically must be considered when plotting location. This factor adds additional complexity to such an analysis. Thankfully for *Equus*, however, it is known to only have existed since the Pliocene (e.g., McKenna and Bell 1997), and, and such, past continental configurations are not of concern in the case of this genus.
4. Knowledge about the full distribution of the genus *Equus* potentially has widespread applications for evolutionary biology in terms of speciation and extinction events leading to extant distributions. The *Equus* use case also pertains to other disciplines including, for example, conservation management and policy and veterinary medicine for an understanding of the spread of disease (e.g., Moehlman 2002).

Methods: Accessing the Relevant Databases

The overarching goal of this use case was to reconstruct the "complete" distribution of *Equus* from the burgeoning set of platforms providing location records and to integrate these across the divide between paleontological and neontological data resources. We did taxonomic search queries using the name "*Equus*" from six databases: iDigBio, Paleobiology Database, VertNet, BISON, Neotoma, and GBIF. We selected these six databases from a priori knowledge of what kinds of data they were likely to contain. We realize that these databases are not all similar in their design, intent, and attributes; for example, they contain different kinds of source data records, that is, some are vouchered specimens, whereas others are nonvouchered occurrences or reports based on the literature. We also understood that whereas some of these databases have unique data occurrences, others overlap with two or more of the databases selected, for example, VertNet data are also published on GBIF.

In this paper we use the term "location" as a separate concept from "occurrence." This convention follows the DwC biodiversity information standards (Taxonomic Database Working Group 2016). The term "dwc:Location" has related data properties such as mean latitude and longitude or, strictly speaking, dwc: decimalLatitude and dwc:decimalLongitude. Here, "dwc" is an abbreviation for the "http://rs.tdwg.org/dwc/terms/," which describes

TABLE 1. Data standards that pertain to the temporal, age, and related geological context of fossil specimens contained in the databases described here.

A. Darwin Core (http://rs.tdwg.org/dwc/terms; iDigBio, VertNet, BISON, GBIF)*

Geological context	
geologicalContextID	earliestAgeOrLowestStage
earliestEonOrLowestEonothem	latestAgeOrHighestStage
latestEonOrHighestEonothem	lowestBiostratigraphicZone
earliestEraOrLowestErathem	highestBiostratigraphicZone
latestEraOrHighestErathem	lithostratigraphicTerms
earliestPeriodOrLowestSystem	group
latestPeriodOrHighestSystem	formation
earliestEpochOrLowestSeries	member
latestEpochOrHighestSeries	bed

B. Paleobiology Database (https://paleobiodb.org/cgi-bin/bridge.pl?a=displaySearchStrataForm)

Stratigraphic search values†
Group, formation, or member
Time interval (or age in Ma)
Paleoenvironment
Lithology

C. Neotoma (http://www.neotomadb.org/uploads/NeotomaManual.pdf)‡

Lithology	**Geochronology**	**Relative age**	**Sample ages**	**Tephrachronology**
LithologyID	AgeTypeID	RelativeAgeID4	SampleAgeID	TephraID
CollectionUnitID	Age	RelativeAgeUnitID	SampleID	TephraName
DepthTop	ErrorOlder	RelativeAgeScaleID	ChronologyID	C14Age
DepthBottom	ErrorYounger	RelativeAge	Age	C14AgeYounger
LowerBoundary	Infinite	C14AgeYounger	AgeYounger	C14AgeOlder
	Delta13C	C14AgeOlder	AgeOlder	CalAge
		CalAgeYounger		CalAgeYounger
		CalAgeOlder		CalAgeOlder

*BISON has a dwc:basisOfRecord field that can be used to search on fossil records, but the corresponding geological context data are not included, e.g., in a .csv file of the extinct taxon *Equus scotti* retrieved on 25 August 2016. GBIF aggregates data that are both DwC and non-DwC compliant.

†Although group, formation, and member are similar to the DwC data standards terminology, those for PBDB are not DwC compliant.

‡These descriptors pertain to the geological, temporal context, and age of the fossil sites included in this database. Because lithology is included in the PBDB, the corresponding lithology standards are included here. Some data standards, e.g., notes, are not included here; thus some of the data standards sets are abridged. Extensive additional documentation is provided in http://www.neotomadb.org/uploads/NeotomaManual.pdf.

the collection of all DwC properties, classes, and encoding schemes.

Most of the databases that combine both fossil and extant data have the ability to identify fossil versus extant via the dwc:basisOfRecord convention. Nevertheless, a fundamental problem that limits the research utility of temporal data for fossils is the fact that there is no corresponding uniform set of DwC conventions that guide temporal data across all relevant platforms (Table 1).

We use the term "extant" to mean species that exist today. This is different from the term Holocene or Recent used by geologists, which is the geological time interval over the past ~10,000 radiocarbon years since the end of the Pleistocene (Gibbard and van Kolfschoten 2004). Extant specimens typically are curated in neontological collections.

Results of Data Exploration and Mining

As summarized in Table 2, the mining of *Equus* data from these sources yielded 123.8 K location records (in reporting records located, K is used throughout for 1000), ranging from 44.5 K (GBIF) to 0.2 K (Neotoma). As a relevant subset of these data, 116.2 K fossil specimens are identified primarily via the dwc:basisOfRecord, ranging from 42.4 K in GBIF to 0.2 K in Neotoma. Of the six databases mined for *Equus*,

TABLE 2. Summary characteristics of the six databases used in this metaresearch study of *Equus*. See text for discussion of iDigBio, PBDB, and GBIF and Supplementary Document 1 for VertNet, BISON, and Neotoma.

Name	Website	Year started	Total no. of data points*	Principal record type†	Primarily Recent (R) or fossil	Data records duplicate	No. of *Equus* records	No. of fossil records retrieved‡	Primary coverage *Equus*	Most recent date retrieved
iDigBio	www.idgbio.com	2012	64.6 M	V	Mostly R	GBIF	22.4 K	21.9 K	North America	9 August 2016
PBDB	https://paleobiology.org	1998	1.3 M	O	Fossil	GBIF	1.6 K	1.6 K	World	27 August 2016
GBIF	www.gbif.org	2001	642 M	Both	Mostly R	various	44.5 K	42.4 K	World	9 August 2016
VertNet	www.vertnet.org	2010	18.7 M	Mostly V	Mostly R	GBIF	29.8 K	25.6 K	World	9 August 2016
BISON	http://bison.usgs.ornl.gov	N.A.	261.7 M	Both	Mostly R	GBIF	25.3 K	24.5 K	United States§	9 August 2016
Neotoma	www.neotomadb.org	2008	Not determined	Varied	Fossil	None	0.2 K	0.2 K	United States	9 February 2016
						Sum	**123.8 K**	**116.2 K**	**Mostly United States, North America**	

*M, million(s); K, thousands(s); as of February 2016.

†V, vouchered specimen in repository; O, occurrence record (not vouchered, or indirectly vouchered).

‡Using DwC BasisofRecord when applicable (iDigBio, GBIF, VertNet, Bison).

§Mandated by BISON mission.

Records include many different kinds of data related to, e.g., vertebrate faunal remains, paleobotany, invertebrates, geochemistry, stratigraphy and geochronology. They do not refer to vouchered museum specimens like some of the other databases.

the three (iDigBio, Paleobiology Database, and GBIF) described below are considered to be primarily useful for the goal of this study. In this particular use case, although each has their own strengths, the other three databases (VertNet, BISON, and Neotoma) are considered to be less useful and these are described in Supplementary Document 1.

Integrated Digitized Biocollections (iDigBio)

The iDigBio portal started in 2012 (Table 2) and as of the middle of 2016 it contains 64.6 million vouchered specimen records (and 14.5 million media records, including photos, 3D images, and sound). There is considerable discussion in the literature about the importance of vouchered natural history specimens (Bradley et al. 2014; McLean et al. 2016); this is one aspect of iDigBio that sets it apart from some of the other databases described here. The iDigBio database aggregates museum, college/university, and related (e.g., herbaria) specimens from more than 250 nonfederal institutions primarily from North America, although some international institutions have also sent their relevant data to iDigBio. (Relevant specimen records in federal repositories are discussed for BISON in Supplementary Document 1.) The locality data are potentially worldwide and initially have focused on extant data records. Paleontological collections (e.g., ANSP, FLMNH, UCMP, YPM) are aggregated into iDigBio, and more are expected as additional initiatives are undertaken in the future (e.g., iDigPaleo and ePANDDA). The iDigBio portal organizes the fossil data using the DwC Geological Context standards (Table 1). Documentation about the iDigBio API can be found on the "Technical Information" page of the iDigBio website (see Appendix) and Page et al. (2015) provide a general overview of this project and the umbrella ADBC program within NSF.

A search on "*Equus*" on the iDigBio portal can be performed in at least two different ways. One is simply on the "Search Records" line, which retrieved 32.8 K records. The problem with this query is that it searches fields other than strictly taxonomic ones and retrieves records for which the string *Equus* occurs. For example, in this search for *Equus*, the genus *Boophilus microplus* (also referred to *Rhipicephalus*), a livestock tick (known to occur on *Equus*), is retrieved at the same time. In contrast, however, entering *Equus* in the dwc: scientificName enabled field, with EOL synonyms selected, yields 22.4 K results (Fig. 1A), including 21.9 K fossil specimens (Fig. 1B). The high percentage (97.7 %) of fossils likely relates to the widespread abundance and large collections of extinct *Equus* relative to the difficulty of conserving large mammal specimens of this genus in most museum collections. With regard to the full distribution in space and time, these results show (Fig. 1A) many location records in North America and, to a lesser extent, South America, Africa, and Europe. Based on prior knowledge (e.g., Nowak 1999), there are some notable absences in this output, for example in Eurasia, where vouchered specimen data records in museum collections are apparently scarce. Likewise, as we will see below from the PBDB, many fossil *Equus* locality records are also poorly represented or lacking on continents where they are known to have existed, likely reflecting the fact that only a few major fossil vertebrate collections outside North America are currently aggregated into iDigBio.

The obvious difference between the two plots in Figure 1 is the increased geographic spread to localities in Africa based on extant specimens (Fig. 1A), as opposed to those represented by fossils. In Figure 1A there is an obvious outlier for an occurrence in the Bight of equatorial Africa in the southern Atlantic Ocean. This represents 12 extant specimens of African *Equus* in a single collection for which there is a common error where coordinates are likely incorrectly listed in the source database as latitude 0, longitude 0.

Paleobiology Database (PBDB)

The PBDB started in 1998 as a community-driven resource of fossil data primarily mined from collections and the literature. As of the middle of 2016, it includes about 1.3 million individual records ("occurrences") ranging in age from the Precambrian to the Quaternary. Its intent is to primarily focus on taxonomic,

FIGURE 1. Plots of location for *Equus* records (retrieved 10 August 2016) using the integrated mapping function of iDigBio. A, All specimens. B, Fossil specimens.

occurrence, location, and geological (temporal) data, although other information types, such as bibliographic references, are also available. PBDB is not DwC consistent, for example, fossil occurrences in PBDB use "occurrence" data entries instead of dwc:decimalLatitude and dwc:decimalLongitude. Likewise, geological and temporal standards are also different from those in DwC (Table 1). Nevertheless, PBDB data can be exported to other databases using DwC standards (PDBD 2016). PBDB does not use data directly connected to

vouchered museum specimens. In a sense, however, many of these records are indirectly vouchered, that is, they could be verified from the original publications from which the data were mined if they are based on catalogued specimens in permanent repositories. PBDB does not typically include extant taxa. Clear and unique strengths of the PBDB include the number of fossil data records and the associated geochronological data. (A previous attempt, the Paleoportal, which no longer is supported, has some similar attributes based on vouchered museum specimens.) Peters and McClennen (2015) provide an overview of the PBDB, including API documentation.

A query for *Equus* in the "Taxonomic Name Search" function in the PBDB yielded 1.6 K occurrence records (Fig. 2). In addition to some overlapping occurrence (location) records, for example, in the United States and sub-Saharan Africa, the PBDB has some differences relative to iDigBio. These are obvious by visual comparison of Figure 1A with Figure 2. For example, the PBDB has occurrence coverage for numerous Quaternary localities in Canada. Likewise, the PBDB increases the number of occurrence records in South America and Africa and adds many others, particularly in eastern Europe and Asia, which include numerous Quaternary localities containing *Equus*. Australia never had native equids (Nowak 1999), and *Equus* was introduced to that continent during historical times; it is therefore not surprising that neither iDigBio nor PBDB produced any locality records on that continent.

Global Biodiversity Information Facility (GBIF)

GBIF was established in 2001 as an international open data repository for biodiversity and related subjects. As of the middle of 2016 GBIF contains 642 million records (Table 2), of which 564 million are georeferenced. These data are aggregated from more than 400 sources ranging from vouchered museum collections to nonvouchered occurrence data collected by expert amateurs, for example, iNaturalist. The intention of GBIF is to serve a variety of research uses, including the study of taxonomic occurrences, but also investigations of invasive species, climate change, conservation, human health, agriculture, ecosystem services, and phylogenetics. Of relevance to the current study, GBIF publishes data from and shares data with iDigBio, PBDB, VertNet, BISON, and Neotoma. Nearly 1700 publications are based on GBIF-derived data

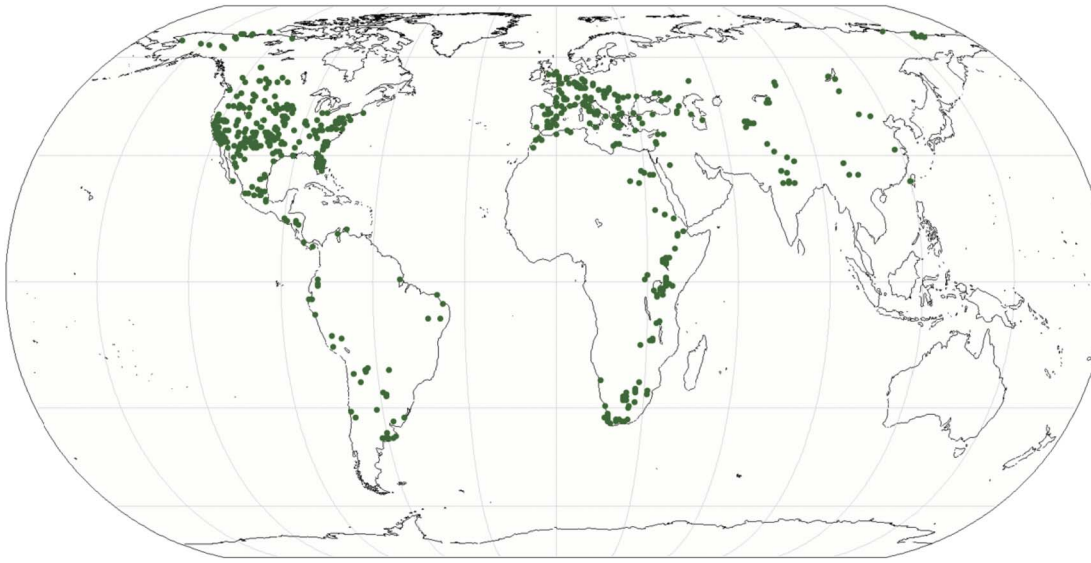

FIGURE 2. Plot of 1.6 K occurrence records of *Equus* using the integrated mapping function of PBDB (retrieved 14 February 2016).

(GBIF 2016), and this number continues to increase each year. GBIF aggregates data that are either DwC (e.g., VertNet) or non-DwC (PBDB) compliant.

An initial search of GBIF in January 2016 using the dwc:scientificName yielded 25.8 K location records. At that time, GBIF did not have an integrated mapping function available, so we downloaded a .csv file and went through a data-quality and cleaning routine to provide a mappable set of records (see discussion in Supplementary Document 2). Nevertheless, a similar query on 10 August 2016 yielded 44.5 K data points for *Equus* (Fig. 3A), including 42.4 K (95.3%) fossil specimens (Fig. 3B). At the present time, GBIF has available an experimental,

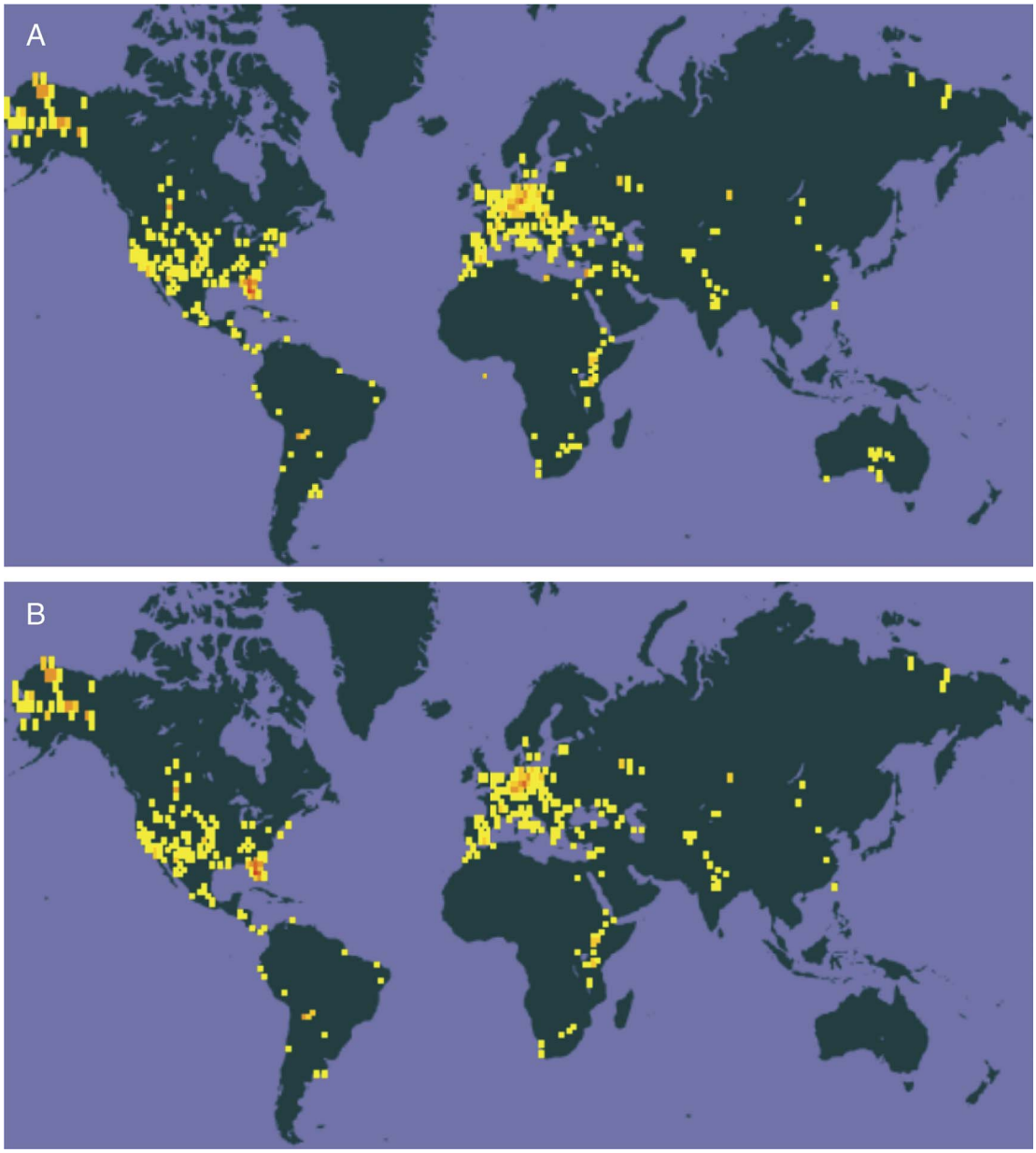

FIGURE 3. Plots of fossil and extant *Equus* location data using the integrated mapping function in GBIF (retrieved 10 August 2016). A, All *Equus*, i.e., extant and fossil. B, Fossil *Equus*.

integrated mapping function that successfully plotted the relevant data points for *Equus* (Fig. 3) without the transposition errors found in the previous .csv file that was downloaded.

GBIF has worldwide coverage, although the data appear to be concentrated in the United States and Europe, with spotty coverage from Canada, South America, Africa, and Asia. Similar to VertNet (Supplementary Document 1), GBIF has extant records for Australia (Fig. 3A). There is surprising overlap between these two plots in Figure 3. Some notable differences include the presence of *Equus* in Australia for all of the data records (including extant), versus the lack of this genus on that continent in the fossil plot. This makes sense, because *Equus* was introduced in Australia as a domesticate during historical times (Nowak 1999). In Figure 3A there is an obvious outlier for an occurrence in the Bight of Africa at the equator in the southern Atlantic Ocean, which almost certainly is a record with geospatial coordinates improperly entered in the source database as latitude 0, longitude 0.

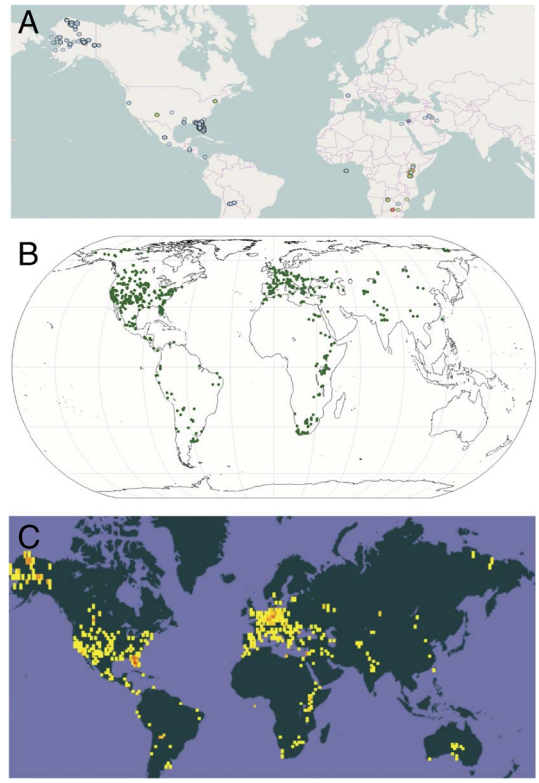

FIGURE 4. Comparison of all *Equus* location data. A, iDigBio (extant and fossil). B, PBDB (fossil). C, GBIF (extant and fossil).

Discussion

General Analysis and Comparison

The exploration and mining of the six databases retrieved 123.8 K locality records for the global distribution of fossil and extant *Equus*. In rank order, the number of data points retrieved from each is as follows: GBIF (44.5 K), VertNet (29.8 K), BISON (25.3 K), iDigBio (22.4 K), PBDB (1.6 K), and Neotoma (0.2 K). Of these, 116.2 K include fossil specimen records of *Equus* (Table 2). Based on previously published references, it is known that extant *Equus* is either native or has been introduced in North America, South America, Europe, Asia, Africa, and Australia. Fossil records of this genus are also previously reported from all of these continents, with the exception of Australia where it was introduced as a domesticate during historical times (McKenna and Bell 1997; Nowak 1999). The location records retrieved from the six databases generally reflect the predicted continental distributions of *Equus*, however, with the possible exception of GBIF for all the data and PBDB for the fossil data, the pattern of coverage is skewed primarily toward North America. There are several reasons for this disparity, likely reflecting the bias in museum vouchered specimens in iDigBio and VertNet and the limited geographic focus of BISON and Neotoma. Figure 4 shows a graphical comparison of the distribution and coverage for all data for iDigBio, PBDB, and GBIF. Of the three, GBIF encompasses the most comprehensive coverage of the biodiversity databases studied here. However, with far fewer records, the PBDB also demonstrates relatively widespread coverage for fossil data.

Where to Find Relevant Data?

We chose to explore six platforms, each with massive quantities of data and all still growing. These selections were based on our prior

knowledge of their general characteristics, for example, museum specimens or other similar data likely recording extant and/or fossil location data where the genus *Equus* is likely to occur or have occurred. For investigators unfamiliar with where to look for relevant biodiversity data, this might present more of a challenge. Some other disciplines have meta-databases such as the National Center for Biotechnology Information, which covers biomedical and genomic research, but biodiversity and paleodiversity databases are not yet as coordinated, centralized, and well linked, and the integration of fossil and extant data remains challenging across platforms. The recently initiated ePANDDA project is directed toward database integration and likely will address the current challenge of exploring and mining data records from relevant sources. Similarly, the GeoDeepDive project (2016) may provide a way in the future for researchers lacking sufficient prior knowledge to know where to look for relevant data.

Data Quality, Integrity, and Pitfalls

Location records mined from large-scale aggregators are fundamental to biodiversity analyses. These data form the basis for first assessments of taxonomic distributions and provide crucial information on species niche profiles in the past and present. When these data are plotted, either from integrated mapping algorithms (e.g., iDigBio, PBDB, BISON, and Neotoma) or by using independent mapping tools (e.g., QGIS) to visualize raw data, an initial means of quality control can be applied through visual inspection for outliers, as we did for iDigBio and GBIF above for the African Bight. As further described in Supplementary Document 2, our initial explorations yielded issues with the data download that were then mitigated more recently by plotting directly from an integrated mapping function in GBIF.

The issue of big biodiversity data quality control and screening (curation) is a much larger problem that has been extensively discussed in the literature for neontological specimens (e.g., Hill et al. 2010, Gaiji et al. 2013; Maldonado et al. 2015) but still has many challenges. Data quality and fitness for use require a multifaceted approach that considers both the spatial and temporal aspects. Recent studies such as Otegui and Guralnick (2016) address semiautomated approaches for flagging problematic records, using a REST API services approach along with an R-package "wrapper" for utilizing that service using standard tools. Likewise, Otegui and Ariño (2012) show the utility of visualization tools that can help detect the shape of data along temporal, spatial, and taxonomic dimensions. When new data-quality services and visualization tools are coupled, end-users can much more quickly find records that are likely to need further data curation. With regard to integrated paleontological and neontological databases, which have yet to be carefully scrutinized, new filters that provide quick ways to sort those records from extant ones using dwc:basisOfRecord will help insure that paleontological data are not simply flagged as outliers. Such efforts rely on data publishers to properly use dwc:basisOfRecord to identify "FossilSpecimen," as opposed to other terms in a recommended (but not fully enforced) controlled vocabulary also including "PreservedSpecimen," "LivingSpecimen," "HumanObservation," and "MachineObservation." It is still unclear whether all such fossil specimen data in repositories such as GBIF are consistently recorded, thus potentially leading to further challenges with truly integrative studies.

Geographic Bias

Our study indicates a considerable bias toward North American records of *Equus* in some databases, for example, BISON because of the U.S-centric mission of the project or Neotoma because of the geography of its data sources. On the other hand, particularly for North America and Europe, PBDB does a better job of less-biased distributions in the fossil record, as does GBIF for combined extinct and extant *Equus* location data. Nevertheless, to a greater or lesser degree, inherent biases in these databases will prevent a truly global understanding of the paleo(biogeography) of the genus *Equus*. Given this inherent bias, more fine-scaled discussions about extinction

geography in the future will be best focused on the North American or European record.

The Holy Grail: Fossil Specimens, Deep Time, and Limitations of Temporal Data

The only place that biodiversity researchers can find answers to questions posed in deep time is via the fossil record. A corollary to this fact is that the ability to better understand macroevolution and extinction depends upon high-resolution data consistently archived across big data platforms, whether or not these include vouchered or other kinds of specimen and data records. One current challenge is the integration of data across platforms when they have different data standards. Of the six databases studied here, four consistently or primarily use DwC; the PBDB does not, but data can be exported to DwC; and Neotoma has such a fundamentally different kind of data structure that cross-platform integration will be a special challenge.

Another challenge is how different geochronological age intervals are named and the complexity of synonymizing these across platforms. For example, in different databases *Equus* specimens may be listed as Rancholabrean in one, Quaternary in another, and late Pleistocene in yet another, or may have associated radiocarbon age determinations. Although these may be geologically contemporaneous, it is impossible to tell for certain with only these geochronological data. The interconnection of these temporal data for research is thus a challenge. Furthermore, for fine-scale evolutionary and extinction events like we might want to explore for *Equus*, these temporal intervals do not provide sufficient resolution and may ultimately require the use of high-precision radiocarbon dates. These kinds of absolute age determinations are only beginning to find their way into big biodiversity databases such as Neotoma. In contrast, the lack of standardized means to report absolute age determinations in DwC is a major drawback to the fundamental importance of these data for paleontological and archaeological data. Given that DwC already maintains a set of terms related to dwc:geologicalContext, the paleontological community would be well served to adopt these standards and thus provide a pathway for better data integration across big data repositories. The next decade will hopefully find major improvements in the interoperability and integration of diverse data within these big biodiversity databases.

Moving Forward: The Future of Big Biodiversity Databases

Equus Extinction Geography

So far as we know, *Equus* became extinct in the Americas about 10,000 years ago during the late Pleistocene great megafaunal extinction (e.g., Koch and Barnosky 2006; Smith et al. 2016). However, the dynamics of how and where this happened are still elusive. With the advent of large databases for fossil records and advances in fields such as niche modeling (e.g., Martinez-Meyer et al. 2004), in theory these geographic patterns of extinction could be elucidated. However, at the present time, neither the relevant location data nor the precise temporal chronology and associated climatic parameters are sufficiently dense and well integrated to resolve this interesting pattern. In this regard Neotoma leads the field with associated radiocarbon age determinations, which exemplify the kind of data needed to study these fine-scale patterns. While location data records are the low-hanging fruit for (paleo)biogeographic analyses, the addition of other kinds of ancillary data represents enormous potential for expanding the kinds of questions that will be answered using big biodiversity databases.

Ancillary Data Fields

As of now, the strength of many of the large databases, including those described above, is in high-quality taxonomic and location data that can be used to answer research questions about (paleo)geography that were previously unknowable. In the current context, ancillary data are other kinds of data that are "attached" to vouchered museum specimen records that ultimately add value to the research capacity of big databases. As mentioned above, the addition of data fields with high-precision temporal age constraints will enhance the spatial and temporal potential of big data

records. These could be used to better understand late Pleistocene megafaunal dynamics and extinction geography, including those of the genus *Equus*. Other data enhancements that are already being attached in some databases such as VertNet and iDigBio include media, such as photos and sounds. With the recent explosion of 3D printing, the application of this technology will further enhance the research capacity of these big databases as 3D images (e.g., MorphoSource, see Appendix) are made available online and either aggregated or linked to big data platforms.

In the last century, and in what now seems to be the Stone Age, MacFadden et al. (1999) spent months acquiring carbon ($\delta^{13}C$) and oxygen ($\delta^{18}O$) stable isotope data from fossil *Equus* teeth. This was done to understand the latitudinal distribution of C_3 and C_4 plants and temperature-dependent data for North and South America during the Pleistocene. A decade from now it is conceivable that a researcher will submit a query to retrieve *Equus* location records that also have stable isotopes as part of diverse trait data, all associated with the vouchered specimens. These could then be plotted instantaneously to display geographic patterns using these kinds of ancillary proxy data. New initiatives such as IsoBank (Pauli et al. 2015), using a model similar to GenBank, have the potential to greatly expand the kinds of interesting taxonomic, geographic, and ecological questions that can be asked by researchers. Moran et al. (2016) have initiated a pilot project in which published, or "legacy," $\delta^{13}C$ and $\delta^{18}O$ stable isotope data are attached to fossil vertebrate specimens in the FLMNH collections via Specify (2016). Plans are in the works for these data to ultimately be aggregated by iDigBio.

Horses in the Cloud: Concluding Comments

The use of "big data" for biodiversity and distributional studies, using either extant and extinct taxa, or both together, has immense potential for asking and answering interesting research questions that were, from an epistemological point of view, previously unknowable. The use of these data sources is not without pitfalls, however, as is exemplified by the use case of *Equus* presented here. One needs to know where to go to look for relevant data, the intentions and data structure of individual big databases, how to screen relevant records for quality control, and, ultimately, how to integrate these data across platforms. It is likely that as the use of big databases continues to increase, further automation such as geospatial cleaning or tools that help with unraveling duplicated records from multiple repositories will become increasingly more valuable for research. Further developments, such as the addition of ancillary data, including images (including both photos and 3D), age determinations (e.g., radiocarbon dating), and ecological information (e.g., stable isotope and phenotypic measurements) will further increase the utility of these platforms in the future. At that point we also will look for further integration of the databases within our content domain and those from other disciplines, such as conservation and veterinary medicine, to enhance our understanding of biodiversity and its ramifications in the twenty-first century.

Acknowledgments

This research, including the development of the iDigBio portal, was supported by National Science Foundation grants EF 1115210 and DBI 1547229. We appreciate discussion about the topic of this paper with D. S. Jones, P. L. Koch, and L. M. Page and helpful input from the anonymous reviewers. This is University of Florida Contribution to Paleobiology number 825.

Literature Cited

Bradley, R. D., L. C. Bradley, H. J. Gardner, and R. J. Baker. 2014. Assessing the value of natural history collections and addressing issues regarding long-term growth and care. BioScience 64: 1150–1158.

Constable, H., R. Guralnick, J. Wieczorek, C. Spencer, and A. T. Peterson. 2010. VertNet: a new model for biodiversity data sharing. PLoS Biology. doi: 10.1371/journal.pbio. 1000309.

Gaiji, G., V. Chavan, A. H. Ariño, J. Otegui, D. Hobern, R. Sood, and E. Robles. 2013. Content assessment of the primary biodiversity data published through the GBIF network: status, challenges and potentials. Biodiversity Informatics 8:94–172.

GeoDeepDive. 2016. i.stanford.edu/hazy/geo, accessed 16 February 2016.

Gibbard, P., and T. van Kolfschoten. 2004. The Pleistocene and Holocene epochs. Pp. 441–452 in F. M. Gradstein, J. G. Ogg, and

A. G. Smith, eds. A geological time scale 2004. Cambridge: Cambridge University Press.
Global Biodiversity Information Facility. 2016. www.gbif.org, accessed 28 January 2016.
Hill, A. W., J. Otegui, A. H. Ariño, and R. P. Guralnick. 2010. GBIF position paper on future directions and recommendations for enhancing fitness-for-use across the GBIF network, Version 1.0. Global Biodiversity Information Facility, Copenhagen. http://www.gbif.org/resource/80623.
Koch, P. L., and A. D. Barnosky. 2006. Late Quaternary extinctions: state of the debate. Annual Review of Ecology, Evolution, and Systematics 37:215–250.
Kousta, S., C. Ferguson, and E. Ganley. 2016. Meta-research: broadening the scope of PLOS Biology. PLoS Biology 14:e1002334.
Linnaeus, C. 1758. Systema naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis. Editio decima, reformata. Holmiae [Stockholm], Laurentii Salvii.
MacFadden, B. J., T. E. Cerling, J. M. Harris, and J. Prado. 1999. Ancient latitudinal gradients of C3/C4 grasses interpreted from stable isotopes of New World Pleistocene horse (*Equus*) teeth. Global Ecology and Biogeography 8:137–149.
Maldonado, C., C. I. Molina, A. Zizka, C. Persson, J. Albán, E. Chilquillo, N. Rønsted, and A. Antonelli. 2015. Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? Global Ecology and Biogeography 24:973–984.
Martínez-Meyer, E., A. Townsend Peterson, and W. W. Hargrove. 2004. Ecological niches as stable distributional constraints on mammal species, with implications for Pleistocene extinctions and climate change projections for biodiversity. Global Ecology and Biogeography 13:305–314.
McKenna, M. C., and S. K. Bell. 1997. Classification of mammals above the species level. Columbia University Press, New York.
McLean, B. S., K. C. Bell, J. L. Dunnum, B. Abrahamson, J. P. Colella, E. R. Deardorff, J. A. Weber, A. K. Jones, F. Salazar-Miralles, and J. A. Cook. 2016. Natural history collections-based research: progress, promise, and best practices. Journal of Mammalogy 97:287–297.
Moehlman, P. D. R. (ed.) 2002. Equids—zebras, asses, and horses: status survey and conservation action plan. IUCN—The World Conservation Union, Gland, Switzerland.
Moran, S. M., R. C. Hulbert, Jr., W. H. Brown, and B. J. MacFadden. 2016. Increasing the research potential of digitized fossils: a pilot study using Specify to attach stable isotope data to vouchered museum specimens. Geological Society of America Abstracts with Programs 48(7). doi: 10.1130/abs/2016AM-280994.
Nowak, R. M. 1999. Walker's mammals of the world. John Hopkins University Press, Baltimore, Md.
Otegui, J., and A. H. Ariño. 2012. BIDDSAT: visualizing the content of biodiversity data publishers in the Global Biodiversity Information Facility network. Bioinformatics 28:2207–2208.
Otegui, J., and R. P. Guralnick. 2016. The geospatial data quality REST API for primary biodiversity data. Bioinformatics 32:1755–1757.
Page, L. M., B. J. MacFadden, J. A. Fortes, P. S. Soltis, and G. Riccardi. 2015. Digitization of biodiversity collections reveals biggest data on biodiversity. BioScience. doi: 10.1093/biosci/biv104.
Paleobiology Database. 2016. PBDB data services: formats and vocabularies. https://paleobiodb.org/data1.1/formats_doc.html, accessed 3 May 2016.
Pauli, J. N., S. S. Steffan, and S. D. Newsome. 2015. It is time for IsoBank? BioScience. First published online 22 January 2015. doi: 10.1093/biosci/biu230.
Peters, S. E., and M. McClennen. 2015. The Paleobiology Database application programming interface. Paleobiology 42:1–7.
Smith, F. A., C. P. Tomé, E. A. E. Smith, S. K. Lyons, S. D. Newsome, and T. W. Stafford. 2016. Unraveling the consequences of the terminal Pleistocene megafauna extinction on mammal community assembly. Ecography 39:1–17.
Specify. 2016. Specify Software Project. http://specifyx.specifysoftware.org, accessed 10 August 2016.
TDWG. 2016. Taxonomic Database Working Group. http://rs.tdwg.org/dwc/terms/guides/rdf, accessed 3 May 2016.
Wieczorek, J., D. Bloom, R. Guralnick, S. Blum, and M. Döring. 2012. Darwin Core: an evolving community-developed biodiversity data standard. PLoS ONE 7:e29715.
Wolstencroft, K., R. Haines, D. Fellows, A. Williams, D. Withers, et al. 2013. Nucleic Acids Research. Web Server Issue 41:557–561.

Appendix

List of abbreviations and acronyms used in this paper, and the URL links, where applicable.

ADBC, Advancing the Digitization of Biological Collections program at the U.S. National Science Foundation, www.nsf.gov

ANSP, Academy of Natural Sciences Philadelphia of Drexel University, www.ansp.org

API, application program interface

Arctos, Arctos Museum Database, http://arctos.database.museum/

BLM, U.S. Bureau of Land Management, http://www.blm.gov/wo/st/en.html

BISON, Biodiversity Information Serving Our Nation, bison.usgs.ornl.gov.csv, comma-separated values data file

DwC, Darwin Core data standards, http://rs.tdwg.org/dwc/ePANDDA, Enhancing Paleontological and Neontological Data Discovery API, https://steppe.org/research/epandda/

FAUNMAP, now part of Neotoma (see below)

FishNet, www.fishnet2.net

FLMNH, Florida Museum of Natural History, University of Florida, www.flmnh.ufl.edu

GBIF, Global Biodiversity Information Facility, www.gbif.org

HerpNET, www.herpnet.org

iDigBio, Integrated Digitized Biocollections, https://www.idgibio.org

iDigPaleo, see https://steppe.org/event/idigpaleo-portal-idigbio-paleo-digitization-working-group-webinar/iNaturalist, http://www.inaturalist.org/

ITIS, Integrated Taxonomic Information System, www.itis.gov

MaNIS, Mammal Networked Information System, http://manisnet.org, subsumed by VertNet

MorphoSource, www.morphosource.org
NMNH (Smithsonian), U.S. National Museum of Natural History, http://naturalhistory.si.edu/
NPS, U.S. National Park Service, https://www.nps.gov/index.htm
ORNIS, www.ornisnet.org, replaced by VertNet
PANGAEA, data publisher for the earth and environmental sciences, https://www.pangaea.de/
PBDB, Paleobiology Database, https://paleobiodb.org; see also http://fossilworks.org/
QGIS, a free and open-source geographic information system, www.qgis.org
REST, Representational State Transfer, a Web service to assess the geospatial quality of primary biodiversity data (Otegui and Guralnick 2016)
TDWG, Taxonomic Database Working Group Biodiversity Information Standards, http://www.tdwg.org/
UCMP, University of California Museum of Paleontology, www.ucmp.berkeley.edu
USDA, U.S. Department of Agriculture, http://www.usda.gov/wps/portal/usda/usdahome
USFS, U.S. Forest Service, http://www.fs.fed.us/
USFWS, U.S. Fish and Wildlife Service, http://www.fws.gov/
USGS, U.S. Geological Survey, https://www.usgs.gov/
VertNet, http://vertnet.org
YPM, Yale Peabody Museum, www.peabody.yale.edu