The increasing disconnection of primary biodiversity data from specimens: How does it happen and how to handle it?

Julien Troudet, Régine Vignes-Lebbe, Philippe Grandcolas, Frédéric Legendre

To cite this version:

Julien Troudet, Régine Vignes-Lebbe, Philippe Grandcolas, Frédéric Legendre. The increasing disconnection of primary biodiversity data from specimens: How does it happen and how to handle it?. Systematic Biology, 2018, 67 (6), pp.1110-1119. 10.1093/sysbio/syy044 . hal-02314020

Running head

SHIFT IN PRIMARY BIODIVERSITY DATA

Title

The increasing disconnection of primary biodiversity data from specimens: How does it happen and how to handle it?

Authors' names

JULIEN TROUDET, REGINE VIGNES-LEBBE, PHILIPPE GRANDCOLAS AND FREDERIC LEGENDRE*

Authors' affiliations

Institut Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, 57 rue Cuvier, CP50, 75005 Paris, France

**Correspondence to be sent to: Institut Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, 57 rue Cuvier, 75005 Paris, France; E-mail:* frederic.legendre@mnhn.fr

Abstract.—Primary biodiversity data represent the fundamental elements of any study in systematics and evolution. They are, however, no longer gathered as they used to be and the mass-production of observation-based occurrences is overthrowing the collection of specimen-based occurrences. Although this change in practice is a major upheaval with significant consequences in the study of biodiversity, it remains understudied and has not attracted yet the attention it deserves. Analyzing 536 million occurrences from the Global Biodiversity Information Facility (GBIF) mediated data, we show that this spectacular change affects the 24 eukaryote taxonomic classes we targeted: from 1970 to 2016 the proportion of occurrences marked as traceable to tangible material (i.e. specimen-based occurrences) fell from 68 to 18 %; moreover, most of those specimen based-occurrences cannot be readily traced back to a specimen because the necessary information is missing. Ethical, practical or legal reasons responsible for this shift are known, and this situation appears unlikely to be reversed. Still, we urge scholars to acknowledge this dramatic change, embrace it and actively deal with it. Specifically, we emphasize why specimen-based occurrences must be gathered, as a warrant to allow both repeating evolutionary studies and conducting rich and diverse investigations. When impossible to secure, voucher specimens must be replaced with observation-based occurrences combined with ancillary data (e.g. pictures, recordings, samples, DNA sequences). Ancillary data are instrumental for the usefulness of biodiversity occurrences and we show that, despite improving technologies to collate them, they remain rarely shared. The consequences of such a change are not yet clear but we advocate collecting material evidence or ancillary data to ensure that primary biodiversity data collected lately do not partly become obsolete when doubtful.

Keywords.—Primary biodiversity data, specimen, observation, database, ancillary data, biodiversity occurrences, big data

Primary biodiversity data, the bricks of systematics and evolutionary studies (May 1990; Funk and Richardson 2002; Hortal et al. 2015), are not gathered nowadays as they used to be. In the early days of systematics, specimens were collected methodically. Today, because of ethical and practical reasons partly imposed by the current biodiversity crisis, unvouchered observation records, i.e. observations with no link to any tangible material, are mainly gathered (Gaiji et al. 2013). Unvouchered observations and vouchered specimens are biodiversity occurrences of different fundamental nature, each having assets and liabilities. Unvouchered observations, for instance, are recorded and shared more rapidly than specimens are collected and databased. With unvouchered observations, biodiversity data accumulate faster than ever (Bisby 2000; Kitchin 2014), but the link to specimens in natural history collections is being lost. We argue here that the change in biodiversity data gathering [from specimen-based (SB) to observation-based (OB) occurrences] has strong consequences in systematics and evolutionary biology. This change must then be analysed, acknowledged, and its effects integrated in our practices; the sooner, the better.

Biodiversity occurrences are not equivalent to one another and, according to their nature (SB or OB, old or recent, with ancillary data or not, etc.), they offer more or less research opportunities (Fig. 1). Generally, a biodiversity occurrence contains a taxonomic identification, a localisation and a date (Ariño 2010). These three pieces of information can be provided for SB or OB occurrences, and, in both cases, can be accurate or not, and more or less precise. Accuracy and precision mostly depend on the collector's skills and equipment, but they are also related to the nature of the primary biodiversity occurrence. In addition, a biodiversity occurrence, be it SB or OB, can be complemented with ancillary data such as pictures or samples, increasing the information content of biodiversity occurrences and their usefulness (Gaiji et al. 2013; Garrouste 2017; Fig. 1). Most ancillary data, however, cannot be gathered *a posteriori* of an OB occurrence, whereas it can be for a SB occurrence. Thus, the way primary biodiversity data are collected impacts their provided information content for current and future investigations.

This change in practice (proportionally much more OB than SB occurrences) is a major upheaval with spectacular consequences for systematics and evolutionary studies. Since the very beginning of systematics, specimens have been collected and used to inventory the diversity of life and later to decipher the relationships within the tree of life (Giribet 2015). Natural history collections (NHC), which now support biodiversity, morphology or molecular databases, have been put together and used for species identification and description, comparative anatomy, and phylogenetic studies, to name a few practices embodying their usefulness (Garner et al. 2014; Kemp 2015; Buerki and Baker 2016; Fig. 1). Databases containing mainly unvouchered observations would not be as profitable as data repositories composed of specimens but they have positive sides in return (e.g. the pace at which biodiversity occurrences are shared; datasets with higher statistical value, etc.) and can be complemented with diverse media. It is pivotal to accommodate to this shift now to make the best of it. As often, good legacy of previous practices and fruitful innovations must be retained and developed, while bad legacy must be put aside (Godfray 2002).

We argue that specimens belong to the good legacy and are too important to be put aside. Even though specimens, like digital data, are not everlasting, their existence offers a high guarantee for repeatability in the study of biodiversity (Huber 1998; Schilthuizen et al. 2015; Turney et al. 2015; Ceríaco et al. 2016; Grandcolas 2017; Gutiérrez and Pine 2017) and the possibility to apply technical advances on them. The recent revolutions in systematics, i.e. the use of DNA and much recently the advent of next generation sequencing (NGS), illustrate this point because they rely on specimens or samples (Pellens et al., 2016; Gutiérrez and Pine 2017). Even better, these technical advances are qualified as revolutionary because specimens are available to use them on, enabling us to engage in new research agenda (e.g. Anmarkrud and Lifjeld 2017). Similarly, in the era of phylogenomics, several authors have recently underlined the necessary revival of morphological studies in systematics, which, again, rely on specimens (e.g. Jenner 2004; Wiens 2004; Smith and Turner 2005; Yassin 2013; Pyron 2015; Wanninger 2015; Wipfler et al. 2016).

Beyond specimens, good practices about items providing additional information content (e.g. samples or pictures) should be advocated to assist the change in biodiversity data gathering (e.g.

Garrouste 2017). Every item of data associated with an occurrence (be either an unvouchered observation or a specimen) is an additional evidence to fight against one or several of the seven currently identified biodiversity shortfalls (Hortal et al. 2015). The Linnean shortfall, the gap between the described species and the actual number of species, undoubtedly requires specimen collection (Ceríaco et al. 2016; Dubois 2017; Pine and Gutiérrez 2018; see Pape et al. 2016 for an opposite opinion). But other shortfalls could be filled, in certain cases, as efficiently with samples or pictures than with specimens. A picture or a DNA sample of a well-known species would efficiently contribute to reduce the Prestonian shortfall, i.e. the lack of knowledge about the abundance of species and their population dynamics in space and time (Cardoso et al. 2011). Indeed, when doubtful, a species attribution can be checked consulting the picture or sequencing DNA, so that observational occurrences with ancillary data constitute appropriate datasets for evolutionary studies.

When a paradigm shift is on the way, measures are required to guide this shift and ensure its maximal usefulness now and in the future. Here, we test whether a shift in the study of biodiversity (i.e. primary data are not SB anymore but mainly OB) is on the rise and whether it is restricted to some organisms. We also investigate when it started, whether it comes with more ancillary and precise data, and how it might affect the fields of systematics and evolution. Analysing 536 million occurrences from the GBIF (Global Biodiversity Information Facility) in 24 taxonomic classes, we show empirically that this shift is widely shared across eukaryotes. From then on, because current decisions will shape the future and to ensure maximal benefits of biodiversity occurrences in systematics and biodiversity research in general, we provide guidelines for primary biodiversity data gathering and sharing, guidelines easily met from individual research to broad citizen science programs.

MATERIALS AND METHODS

Data set

We downloaded all the data available from the GBIF portal in June 2016 (http://doi.org/10.15468/dl.hqesx6). These 649 million occurrences were saved as a Darwin Core archive (www.tdwg.org). Occurrences from this archive were extracted and imported into a SQL database, where they were indexed to reduce computation time of later queries. We focused on 24 taxonomic classes out of the 297 referenced in the GBIF, excluding the classes with less than 1 million occurrences (9.4 million occurrences, distributed into 19 thousands species, had no class affiliation). This filtering reduced the dataset to 626 million of occurrences (NBocc) and 1.01 million species, representing more than 96 % of the total number of occurrences and 84 % of the total number of species in the GBIF. Finally, because we computed statistics over time, data without a year of collection were excluded. We ended up with 536 million occurrences, which is the dataset used to compute all statistics. A lag exists between an occurrence event recording and its integration in the GBIF database (S. Gaiji comm. pers.) and it might be related to the type of occurrences (i.e. specimen- or observation-based). Consequently, even though we show results until 2016, we avoid interpreting the last ten years results in the plots to limit the risk of hazardous conclusions.

Data Quantity

To calculate data quantity in the GBIF mediated data, the number of occurrences collected per year was counted. Then, a data accumulation curve was computed.

Data Origin

In the GBIF, the origin of an occurrence can be specified using a controlled vocabulary in the '*basisOfRecord*' field. As in Troudet et al. (2017), we distinguished "specimen-based occurrences" linked to tangible material from "observation-based occurrences" (or disconnected observations). The category "specimen" regrouped fossil specimen, living specimen, material sample, and preserved specimen. The category "observation" regrouped human observation, machine observation, observation, and literature. Literature occurrences could have been placed in the specimen category, a solution we have discarded because the link to specimen is not straightforward. This choice does not affect the conclusions drawn here because only 497,231 occurrences (i.e. <0.1 %) have a literature origin. A third category, corresponding to the option "unknown", was also kept.

Supporting Files

Supporting files (or links leading to such files) can be associated to an occurrence in the GBIF. They contribute to improve the traceability between a taxon name and a given occurrence. Two kinds of supporting files are mainly used: DNA sequences and multimedia files. For each of those supporting data, we computed 1) the quantity of both DNA sequences and multimedia files per year, and 2) the yearly numbers of DNA sequences and multimedia files divided by the yearly number of occurrences. This last number approximates (because a same occurrence can have several supporting files) the proportion of occurrence with supporting files.

To further understand the structure of the GBIF mediated data we also classified occurrences with supporting files according to their origin (i.e. '*basisOfRecord*'). Thus, we distinguished the number of specimen-based occurrences with multimedia supporting files from the observation-based and unknown occurrences with multimedia supporting files.

Development of Data Completeness

Primary biodiversity data are all the more useful as they are associated to a lot of information. The DarwinCore format currently in use in the GBIF (Wieczorek et al. 2012) provides 234 columns to record information as diverse as the ethology of a living specimen or the geological strata of a fossil specimen. A complete occurrence would never require these 234 columns to be filled, because there are always inapplicable columns for a given occurrence. Nevertheless, the development of data completeness over time can be estimated from the evolution of the proportion of columns containing information. We thus averaged the proportion of non-null (non-empty) columns per occurrence per year.

Development of Taxonomic and Spatial Precision

In general, a primary biodiversity occurrence is associated to a scientific name, which can be more or less precise depending on the skills of the identifier but also on the state and availability of taxonomic knowledge. We estimated taxonomic precision (in number and proportion per year) differentiating occurrences identified at least at the species level from supra-specific occurrences. The proportion of occurrences identified at the species or infraspecific level was used to estimate the taxonomic precision of the GBIF mediated occurrences. As for the development of spatial imprecision, it was calculated as the number and proportion, per year, of occurrences lacking coordinates or flagged in the GBIF as data with coordinate issues.

RESULTS AND DISCUSSION

A Shift in the Recording of Primary Biodiversity Data

In the current context of biodiversity crisis, numerous pleas have incited the scientific community to collect as much biodiversity data as possible, out of the fear it might disappear before we even knew of its existence (May 2004; Butchart et al. 2010). These calls have been heard and, indisputably, biodiversity data accumulate faster than ever (Fig. 2 and Supporting Information), a trend most classes of organisms exhibit even though for a few of them the trend is not so strong (Troudet et al. 2017). The >57 million occurrences submitted to the GBIF in 2014, more than five times the amount of data submitted ten years earlier (i.e. 11 million occurrences in 2004), embody this report (Supporting Information). With this spectacular acceleration, the amount of data available to scientists is so huge that the study of biodiversity has entered into the “Big Data” era (Hampton et al. 2013; Joppa et al. 2016; Kelling et al. 2009). Arguably, this trend reflects the advent of new scientific communities taking or renewing their interest in biodiversity. It may also suggest an increasing appeal of the public for biodiversity. In both cases, this situation offers new opportunities that will enrich our understanding of biodiversity and generate a higher awareness of conservation issues or of biodiversity shortfalls (Hortal et al. 2015). Other benefits followed such as an increased power in statistical analyses because of larger datasets or the possibility to tackle issues at large taxonomical, temporal or spatial scales (Rosenheim and Gratton 2017). However, the large volume of data is also a curation challenge that must be handled to avoid passing on a dubious source of knowledge to future generations because of a fall in data quality (Howe et al. 2008), a criticism regularly brought up for GBIF mediated data (e.g. Yesson et al. 2007).

This acceleration is triggered, at least partly, by a change in the way biodiversity data are recorded. The origin of biodiversity data has shifted from a majority of specimen-based (SB) to a majority of observation-based (OB) occurrences. This shift has been previously suspected (Gaiji et al. 2013) and we show here that, from 1970 to 2016, the proportion of occurrences traceable to tangible material (i.e. specimens) fell from 68 % to 18 %. This decrease is not due to recent digitization initiatives because we used the collection/observation date, and not the digitization date, to compute these statistics. This result applies to the 24 classes studied, except for a few eccentric cases such as Globothalamea and Polychaeta (Figs. 2 and 3). Likely, these exceptions relate to specific practices for observing, collecting or curating these organisms, or to their low volume of primary biodiversity data, which might cast doubt on their atypical trends. Besides, this shift might be slightly inflated because it presumably requires less time to integrate OB than SB occurrences in the GBIF. Still, ignoring the last ten years to limit this potential bias (shaded area in Fig. 2), this shift remains striking. It started, for most of the organisms, in the second half of the 20th century and kept intensifying ever since. On the opposite, the number of SB occurrences has stagnated, or increased marginally at best, in the past 40 years. More worrying, most of SB occurrences cannot be readily traced back to a specimen: Only 238 000 occurrences have a filled “*materialsampleid*” column, representing only 0.28 % of the 84 million SB occurrences. This number illustrates that the way SB occurrences are recorded in biodiversity databases must be improved. Even though a specimen exists somewhere, it cannot be located without a potentially complex investigation procedure. This situation hampers the verification process, a founding step in scientific practice (Turney et al. 2015). Although scientists can be delighted with the pace at which biodiversity data accumulate, they cannot be satisfied with a biodiversity research relying mainly on unverifiable or hardly verifiable occurrences.

Divergent causes, not necessarily exclusive, may explain this practice shift. In a context of massive biodiversity loss, a sense of urgency fueled the pleas for accelerated data collection (Hampton et al. 2013) and encouraged the accumulation of unvouchered observations, less destructive and easier to produce, share and store than specimen-based occurrences. Ethical considerations and conservation issues that hinder specimen collections have commonly been put forward (e.g. Minteer et al. 2014), although they are debatable in some situations (Dubois and Nemésio 2007; Dubois 2009; Dubois 2017; Löbl 2017). The adoption of the Nagoya protocol by many countries and the development of mobile applications have undoubtedly contributed to this shift as well. Concurrently, Grandcolas (2017) suggested that this shift started when biodiversity sciences merged with general biology, more interested in discovering general patterns and laws than in documenting diversity that is supposedly already known enough. Others underlined the lack of human and economic resources to ensure both the gathering of specimens and the curation of natural history collections (Kemp 2015). These reasons could have favoured a decrease in specimen collection. On the other hand, the number of observation-based occurrences has dramatically increased with, for instance, the rise of citizen science that enable to rapidly produce a vast amount of observational data (Dickinson et al. 2012) and that will certainly become more attracting and rewarding for the public in the future (e.g. Silvertown et al. 2015). Given the multiple origins of this trend, it seems unlikely to be reversed in the near future and must be organised and guided to ensure maximal benefits for the study of biodiversity.

Primary Biodiversity Data for systematics and evolutionary studies in the 21st Century: Are We There Yet?

The importance of collecting specimens in taxonomy, evolution and ecology cannot be overemphasized (Huber 1998; Schilthuizen et al. 2015) and two main points, previously discussed in the literature, must be reiterated. First, specimens are needed for species description and for the study of biodiversity in general (Krell and Wheeler 2014; Rocha et al. 2014; Ceríaco et al. 2016; Dubois 2017; Gutiérrez and Pine 2017; Pine and Gutiérrez 2018 *contra* Minteer et al. 2014; Marshall and Evenhuis 2015; Pape et al. 2016). A crucial argument is the utility of specimens for checking species identification. The spectacular growth in biodiversity occurrences is a fantastic opportunity for researchers if, and only if, occurrence quality can be somehow evaluated. Goodwin et al. (2015) assessed that up to half of tropical plant identifications in museum collections were false. Correcting identification errors can be done after examining specimens, but is impossible for unvouchered observations. If Goodwin et al.’s estimation is correct and generalizable to most primary data, the need for some specimens, or at least ancillary data to observation occurrences, is critical. Encouragingly, Kosmala et al (2016) showed that high quality data were obtained in diverse citizen science programs, suggesting that biodiversity data gathered for well-known taxa (Troudet et al. 2017) or geographic areas (Meyer et al. 2016) might contain fewer errors than in the study of Goodwin et al. (2015). Second, the revived focus on morphology advocated lately in systematics requires specimens (Jenner 2004, Wiens 2004, Smith and Turner 2005, Yassin 2013, Pyron 2015, Wanninger 2015, Wipfler et al. 2016). Authors recommending this revival underlined that comparative morphology not only brings phylogenetic characters but also allows including fossil taxa in phylogenetic analyses (e.g. Pyron 2011; Wood et al. 2013), enabling us to better estimate the structure and branch length of the reconstructed trees (Wiens *et al.* 2010; Pyron 2015). Given that phylogenetic thinking has become of paramount importance in biology, improvements in phylogenetic estimation offer large potentialities in comparative analyses and evolutionary studies, and in the study of biodiversity in general (Losos et al. 2013; Buerki et al. 2015).

To happen, this encouraging prospect must be supported with the adequate facilities and workforce to host, curate, describe and identify these specimens. Worryingly, many institutions devoted to these tasks face budget cuts (Kemp 2015). Museums and curators can neither handle the large amount of specimens collected over the world nor ensure the best preservation conditions of these specimens and their identification. It is a critical topic to urgently consider together with this paradigm shift (Kemp 2015; Schilthuizen et al. 2015), and unvouchered observations complemented with ancillary data can contribute to limit this issue.

A specimen is not always necessary for a primary biodiversity data to be useful. Instead of specimens, and in complement to unvouchered observations, digital data or molecular data can be collated. New technologies offer a wide range of tools and methods to collect concrete specimen evidence in nature, and it is now relatively easy and affordable to obtain DNA sequences, images and sound recordings. Then, using molecular and digital data should now be a common practice in the study of biodiversity, as the exponential growth of molecular data and phylogenies, and the development of morphological databases and ontogenies would suggest (Lathe et al. 2008; Parr et al. 2012; Deans et al. 2012, 2015). We show here that digital and DNA data are increasingly used (63,271 and 878,308 ancillary data were collected in 1950 and 2010, respectively) but these data remain patently underemployed (Fig. 4). Only 2.5 % of all the GBIF-mediated occurrences for the 24 focal classes were linked to digital data and 1.5 % to DNA sequences. Worse, proportionally, they become more and more negligible, regarding the large quantity of observations without supporting data. This situation might be improving lately, but the post-2008 tendency observed demands to be confirmed in future years (Fig. 4), which would be happening only if scholars take up the pedagogic and practical challenge of highlighting the importance of ancillary data for biodiversity occurrences. Moreover, and quite inconsistently, digital and DNA data were less used for OB than for SB occurrences (Fig. 5). They would yet be more useful for OB biodiversity data given that they would constitute the only way to independently check or update observation occurrences, whereas one can refer to specimens, as long as those are kept and the traceability chain is not broken, for SB occurrences (Page 2015; Nualart et al. 2017). The high proportion of sequences associated to primary biodiversity data of unknown origin could suggest that when a sample is performed, occurrences are often classified in the catch-all class ‘unknown origin’.

In addition to ancillary data, the usefulness of primary biodiversity occurrences can be maximized through a higher level of precision and completeness in recordings. We expect biodiversity data occurrences to be more precise and complete now than before because tools that are more efficient have been developed. Whatever the nature of the occurrence, spatial coordinates for instance can be easily provided with a high precision level given the democratization of GPS. Data completeness should also improve because of the growing awareness that a global and comprehensive picture of biodiversity is needed. Our results showed that, in proportion, data precision does improve but that data completeness stagnates (Fig. 6 and Supporting Information). The proportion of data with geospatial issues in the GBIF (i.e. data with low spatial precision) decreased from 50.2 % in 1900 to 0.6 % in 2014 in spite of a larger number of occurrences with spatial imprecision – this number being quite stable over the past 30 years (Fig. 6A). Over the same period, records identified at the species level augmented from 89.6% to 99.4%, with once again an increase of supra-species records (Fig. 6B). While species identification and spatial precision improves, so does niche modelling results for instance, which promises significant advances in biogeography (e.g. Meseguer et al. 2015; Töpel et al.

2017). In this regard, important gains for systematics and evolutionary studies can be anticipated from the increasing level of precision in primary biodiversity data.

Given the progresses of technology and the proportion of people owning smartphones with photo and GPS capabilities, targeting a higher level of completeness in biodiversity data is legitimate but the reasons and the necessity of this objective must be well-advertised, a task that falls to scholars. They have the power to modulate the current trend, demanding a minimal amount of ancillary data when designing their personal or collaborative research projects, including citizen science programs. Taking pictures or samples, not necessarily systematically but more often than now, should be part of the scientific protocol. This will not replace the wealth that specimens in natural history collection offer (Funk and Richardson 2002; Buerki and Baker 2016) but would limit the risk that entire datasets become useless when data inaccuracy is suspected. In a recent study, Silvertown et al. (2015) show, for instance, that accurate species identification was achieved using pictures of organisms found in the UK. Whatever its nature and quantity of ancillary data, primary biodiversity data must be made available, and this evolution would require the adequate infrastructures to support the massive amount of data one can foresee. Several data storage and compression options are currently investigated (e.g. Marx 2013; Numanagic et al. 2016), which suggests it will not be an insurmountable hurdle. The costs that should be deployed are substantial but are worth it for evolutionary biologists and for the society. Besides, these efforts would result in large image and DNA databases, whose usefulness, accuracy and automatic search efficiency would augment together with their supply, as a virtuous circle.

The fear of biodiversity disappearance has triggered a vague of biodiversity data accumulation. We are in the middle of a paradigm change where most biodiversity data are not anymore gathered like it used to be. This paradigm change has been undergone without any supervision. Even though some aspects of these changes are highly beneficial, others are suboptimal and must not be ignored. We must act now to allow a better monitoring of the biodiversity research agenda and to continue shaping how biodiversity data should be gathered, diversifying the objects of collection (e.g. specimens, samples, DNA, images, etc. – Knapp 2015). We argue that ancillary data (samples, DNA, pictures) must be collected more methodically than today (Joppa et al. 2016), to avoid disillusionment when we will realize that unvouchered observations were not sufficient to address some current and future preoccupying issues about systematics and evolutionary studies.

ACKNOWLEDGMENTS

This study was developed as part of a Ph.D. project and was funded as a grant by the Ministère de la Recherche to JT. We would like to thank Mark Judson for his comments on an early draft of this manuscript, as well as Marianne Elias and Colin Fontaine for commenting on the latest version. We thank Anne-Sophie Archambeau, Samy Gaiji, Marie-Elise Lecoq, Sophie Pamerlon, Roseli Pellens, Tim Robertson, Dmitri Schigel, Jérôme Sueur and Wilfried Thuillier for fruitful discussions. We thank Alexandre Antonelli and three anonymous reviewers for their constructive comments that contributed to clarify and improve this manuscript.

REFERENCES

Anmarkrud J.A., Lifjeld J.T. 2017. Complete mitochondrial genomes of eleven extinct or possibly extinct bird species. Mol. Ecol. Resour. 17:334–341.

Ariño A.H. 2010. Approaches to estimating the universe of natural history collections data. Biodiv. Inf. 7:81–92

Bisby F.A. 2000. The Quiet Revolution: Biodiversity Informatics and the Internet. Science 289:2309–2312.

Buerki S., Baker W.J. 2016. Collections-based research in the genomic era. Biol. J. Linn. Soc. 117:5–10.

Buerki S., Callmander M.W., Bachman S., Moat J., Labat J.-N., Forest F. 2015. Incorporating evolutionary history into conservation planning in biodiversity hotspots. Phil. Trans. R. Soc. B. 370:20140014.

Butchart SHM, Walpole M, Collen B, van Strien A, Scharlemann JPW, Almond REA, Baillie JEM, Bomhard B, Brown C, Bruno J, Carpenter KE, Carr GM, Chanson J, Chenery AM, Csirke J, Davidson NC, Dentener F, Foster M, Galli A, Galloway JN, Genovesi P, Gregory RD, Hockings M, Kapos V, Lamarque J-F, Leverington F, Loh J, McGeoch MA, McRae L, Minasyan A, Morcillo MH, Oldfield TEE, Pauly D, Quader S, Revenga C, Sauer JR, Skolnik B, Spear D, Stanwell-Smith D, Stuart SN, Symes A, Tierney M, Tyrrell TD, Vié J-C, Watson R. 2010. Global Biodiversity: Indicators of Recent Declines. Science (80-) 328:1164 LP-1168.

Cardoso P., Erwin T.L., Borges P.A.V., New T.R. 2011. The seven impediments in invertebrate conservation and how to overcome them. Biol. Conserv. 144:2647–2655.

Ceríaco, L.M.P., Gutiérrez, E.E., Dubois, A. et al. 2016. Photography-Based Taxonomy Is Inadequate, Unnecessary, and Potentially Harmful for Biological Sciences. Zootaxa 4196:435–45.

Deans A.R., Lewis S.E., Huala E., Anzaldo S.S., Ashburner M., Balhoff J.P., Blackburn D.C., Blake J.A., Burleigh J.G., Chanet B., Cooper L.D., Courtot M., Csösz S., Cui H., Dahdul W., Das S., Dececchi T.A., Dettai A., Diogo R., Druzinsky R.E., Dumontier M., Franz N.M., Friedrich F., Gkoutos G.V., Haendel M., Harmon L.J., Hayamizu T.F., He Y., Hines H.M., Ibrahim N., Jackson L.M., Jaiswal P., James-Zorn C., Köhler S., Lecointre G., Lapp H., Lawrence C.J., Novère N.L., Lundberg J.G., Macklin J., Mast A.R., Midford P.E., Mikó I., Mungall C.J., Oellrich A., Osumi-Sutherland D., Parkinson H., Ramírez M.J., Richter S., Robinson P.N., Ruttenberg A., Schulz K.S., Segerdell E., Seltmann

K.C., Sharkey M.J., Smith A.D., Smith B., Specht C.D., Squires R.B., Thacker R.W., Thessen A., Fernandez-Triana J., Vihinen M., Vize P.D., Vogt L., Wall C.E., Walls R.L., Westerfeld M., Wharton R.A., Wirkner C.S., Woolley J.B., Yoder M.J., Zorn A.M., Mabee P. 2015. Finding Our Way through Phenotypes. PLOS Biol. 13:e1002033.

Deans A.R., Yoder M.J., Balhoff J.P. 2012. Time to change how we describe biodiversity. Trends Ecol. Evol. 27:78–84.

Dickinson J.L., Shirk J., Bonter D., Bonney R., Crain R.L., Martin J., Phillips T., Purcell K. 2012. The current state of citizen science as a tool for ecological research and public engagement. Front. Ecol. Environ. 10:291–297.

Dubois A. 2009. Endangered species and endangered knowledge. Zootaxa 2201:26–29.

Dubois A. 2017. The need for reference specimens in zoological taxonomy and nomenclature. Bionomina 12:4–38.

Dubois A., Nemésio A. 2007. Does Nomenclatural Availability of Nomina of New Species or Subspecies Require the Deposition of Vouchers in Collections? Zootaxa 1409:1–22.

Funk V.A., Richardson K.S. 2002. Systematic Data in Biodiversity Studies: Use It or Lose It. Syst. Biol. 51:303–316.

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

Garner J.L., Amano T., Sutherland W.J., Joseph L., Peters A. 2014. Are natural history collections coming to an end as time-series? Front. Ecol. Environ. 12:436–438.

Garrouste R. 2017. The “wild shot”: photography for more biology in natural history collections, not for replacing vouchers. Zootaxa 4269:453–454.

Giribet G. 2016. New animal phylogeny: future challenges for animal phylogeny in the age of phylogenomics. Org. Divers. Evol. 16:419–426.

Godfray H.C.J. 2002. Challenges for taxonomy. Nature 417:17–19.

Goodwin Z.A., Harris D.J., Filer D., Wood J.R.I., Scotland R.W. 2015. Widespread mistaken identity in tropical plant collections. Curr. Biol. 25:R1066–R1067.

Grandcolas P. 2017. Loosing the connection between the observation and the specimen: a by-product of the digital era or a trend inherited from general biology? Bionomina 12:57–62.

Gutiérrez, E.E., Pine, R.H. 2017. Specimen collection crucial to taxonomy. Science 6331: 1275–1275.

Hampton S.E., Strasser C.A., Tewksbury J.J., Gram W.K., Budden A.E., Batcheller A.L., Duke C.S., Porter J.H. 2013. Big data and the future of ecology. Front. Ecol. Environ. 11:156–162.

Hortal J., Bello F. de, Diniz-Filho J.A.F., Lewinsohn T.M., Lobo J.M., Ladle R.J. 2015. Seven shortfalls that beset large-scale knowledge of biodiversity. Annu. Rev. Ecol. Evol. Syst. 46:523–549.

Howe D., Costanzo M., Fey P., Gojobori T., Hannick L., Hide W., Hill D.P., Kania R., Schaeffer M., St Pierre S., Twigger S., White O., Yon Rhee S. 2008. Big data: The future of biocuration. Nature 455:47–50.

Huber J.T. 1998. The importance of voucher specimens, with practical guidelines for preserving specimens of the major invertebrate phyla for identification. J. Nat. Hist. 32:367–385.

Jenner R.A., Steel M. 2004. Accepting Partnership by Submission? Morphological Phylogenetics in a Molecular Millennium. Syst. Biol. 53:333–359.

Joppa L., O'Connor B., Visconti P., Smith C., Geldmann J., Hoffmann M., Watson J.E., Butchart S.H., Virah-Sawmy M., Halpern B.S. 2016. Filling in biodiversity threats gaps. Science 352:416–418.

Kelling S., Hochachka W.M., Fink D., Riedewald M., Caruana R., Ballard G., Hooker G. 2009. Data-Intensive Science: A New Paradigm for Biodiversity Studies. BioScience 59:613–620.

Kemp C. 2015. The endangered dead. Nature 518:292–294.

Kitchin R. 2014. Big Data, new epistemologies and paradigm shifts. Big Data & Society. 1:2053951714528481.

Knapp S. 2015. The changing role of collections and field research. In: Watson M.F., Lyal C.H.C., Pendry C.A., editors. Descriptive taxonomy: The foundation of biodiversity research. Cambridge University Press, Cambridge: p. 181–189.

Kosmala M., Wiggins A., Swanson A., Simmons B. 2016. Assessing Data Quality in Citizen Science. Front. Ecol. Environ. 14: 551–60.

Krell F.T., Wheeler Q.D. 2014. Specimen collection: Plan for the future. Science 344:815–816.

Lathe W., Williams J., Mangan M., Karolchik D. 2008. Genomic data resources: challenges and promises. Nat. Educ. 1:2.

Löbl I. 2017. Assessing biodiversity: a pain in the neck. Bionomina 12:39–43.

Losos J.B., Arnold S.J., Bejerano G., Iii E.D.B., Hibbett D., Hoekstra H.E., Mindell D.P., Monteiro A., Moritz C., Orr H.A., Petrov D.A., Renner S.S., Ricklefs R.E., Soltis P.S., Turner T.L. 2013. Evolutionary Biology for the 21st Century. PLoS Biol. 11:e1001466.

Marshall S.A., Evenhuis N.L. 2015. New species without dead bodies: a case for photo-based descriptions, illustrated by a striking new species of *Marleyimyia* Hesse (Diptera, Bombyliidae) from South Africa. Zookeys 525:117–127.

May R.M. 1990. Taxonomy as destiny. Nature 347:129–130.

May R.M. 2004. Tomorrow's taxonomy: collecting new species in the field will remain the rate–limiting step. Philos. T. Roy. Soc. B. 359:733–734.

Marx V. 2013. Biology: The big challenges of big data. Nature 498:255–260.

Meseguer A.S., Lobo J.M., Ree R., Beerling D.J., Sanmartín I. 2015. Integrating Fossils, Phylogenies, and Niche Models into Biogeography to Reveal Ancient Evolutionary History: The Case of *Hypericum* (Hypericaceae). Syst. Biol. 64:215–232.

Meyer C., Weigelt P., Kreft H., Lambers J.H.R. 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. Ecol. Lett. 19:992–1006.

Minteer B.A., Collins J.P., Love K.E., Puschendorf R. 2014. Avoiding (Re)extinction. Science 344:260–261.

Nualart N., Ibáñez N., Soriano I., López-Pujol J. 2017. Assessing the Relevance of Herbarium Collections as Tools for Conservation Biology. Bot. Rev. 83:303–325.

Numanagić I., Bonfield J.K., Hach F., Voges J., Ostermann J., Alberti C., Mattavelli M., Sahinalp S.C. 2016. Comparison of high-throughput sequencing data compression tools. Nat. Methods. 13:1005–1008.

Page R.D.M. 2016. DNA barcoding and taxonomy: dark taxa and dark texts. Phil. Trans. R. Soc. B. 371:20150334.

Pape T., 34 signatories. 2016. Taxonomy: Species can be named from photos. Nature 537:307.

Parr C.S., Guralnick R., Cellinese N., Page R.D.M. 2012. Evolutionary informatics: unifying knowledge about the diversity of life. Trends Ecol. Evol. 27:94–103.

Pellens, R., Faith, D.P., Grandcolas, P. 2016. The Future of phylogenetic systematics in conservation biology: Linking biodiversity and society. In: Pellens, R., Grandcolas, P., editors. Biodiversity conservation and phylogenetic systematics: preserving our evolutionary heritage in an extinction crisis. Springer Open, Switzerland : pp. 375-383.

Pine, R.H., Gutiérrez, E.E. 2018. What Is an ‘Extant’ Type Specimen? Problems Arising from Naming Mammalian Species-Group Taxa without Preserved Types. Mammal Review 48:12–23.

Pyron R.A. 2011. Divergence Time Estimation Using Fossils as Terminal Taxa and the Origins of Lissamphibia. Syst. Biol. 60:466–481.

Pyron R.A. 2015. Post-molecular systematics and the future of phylogenetics. Trends Ecol. Evol. 30:384–389.

Rocha L.A., Aleixo A., Allen G., Almeda F., Baldwin C.C., Barclay M.V.L., Bates J.M., Bauer A.M., Benzoni F., Berns C.M., Berumen M.L., Blackburn D.C., Blum S., Bolaños F., Bowie R.C.K., Britz R., Brown R.M., Cadena C.D., Carpenter K., Ceríaco L.M., Chakrabarty P., Chaves G., Choat J.H., Clements K.D., Collette B.B., Collins A., Coyne J., Cracraft J., Daniel T., de Carvalho M.R., de Queiroz K., Di Dario F., Drewes R., Dumbacher J.P., Engilis A., Erdmann M. V, Eschmeyer W., Feldman C.R., Fisher B.L., Fjeldså J., Fritsch P.W., Fuchs J., Getahun A., Gill A., Gomon M., Gosliner T., Graves G.R., Griswold C.E., Guralnick R., Hartel K., Helgen K.M., Ho H., Iskandar D.T., Iwamoto T., Jaafar Z., James H.F., Johnson D., Kavanaugh D., Knowlton N., Lacey E., Larson H.K., Last P., Leis J.M., Lessios H., Liebherr J., Lowman M., Mahler D.L., Mamonekene V., Matsuura K., Mayer G.C., Mays H., McCosker J., McDiarmid R.W., McGuire J., Miller M.J., Mooi R., Mooi R.D., Moritz C., Myers P., Nachman M.W., Nussbaum R.A., Foighil D.Ó., Parenti L.R., Parham J.F., Paul E., Paulay G., Pérez-Emán J., Pérez-Matus A., Poe S., Pogonoski J., Rabosky D.L., Randall J.E., Reimer J.D., Robertson D.R., Rödel M.-O., Rodrigues M.T., Roopnarine P., Rüber L., Ryan M.J., Sheldon F., Shinohara G., Short A., Simison W.B., Smith-Vaniz W.F., Springer V.G., Stiassny M., Tello J.G., Thompson C.W., Trnski T., Tucker P., Valqui T., Vecchione M., Verheyen E., Wainwright P.C., Wheeler T.A., White W.T., Will K., Williams J.T., Williams G., Wilson E.O., Winker K., Winterbottom R., Witt C.C. 2014. Specimen collection: An essential tool. Science 344:814 LP-815.

Rosenheim J.A., Gratton C. 2017. Ecoinformatics (Big Data) for Agricultural Entomology: Pitfalls, Progress, and Promise. Annu. Rev. Entomol. 62:399–417.

Schilthuizen M., Vairappan C.S., Slade E.M., Mann D.J., Miller J.A. 2015. Specimens as primary data: museums and ‘open science’. Trends Ecol. Evol. 30:237–238.

Silvertown J., Harvey M., Greenwood R., Dodd M., Rosewell J., Rebelo T., Ansine J., McConway K. 2015. Crowdsourcing the identification of organisms: A case-study of iSpot. Zookeys. 480:125–146.

Smith N.D., Turner A.H., Macleod N. 2005. Morphology's Role in Phylogeny Reconstruction: Perspectives from Paleontology. Syst. Biol. 54:166–173.

Töpel M., Zizka A., Calió M.F., Scharn R., Silvestro D., Antonelli A. 2017. SpeciesGeoCoder: Fast Categorization of Species Occurrences for Analyses of Biodiversity, Biogeography, Ecology, and Evolution. Syst. Biol. 66:145–151.

Troudet J., Grandcolas P., Blin A., Vignes-Lebbe R., Legendre F. 2017. Taxonomic bias in biodiversity data and societal preferences. Sci. Rep. 7:9132.

Turney S., Cameron E.R., Cloutier C.A., Buddle C.M. 2015. Non-repeatable science: assessing the frequency of voucher specimen deposition reveals that most arthropod research cannot be verified. PeerJ 3:e1168.

Wanninger A. 2015. Morphology is dead - long live morphology! Integrating MorphoEvoDevo into molecular EvoDevo and phylogenomics. Front. Ecol. Evol. 3:54.

Wieczorek J., Bloom D., Guralnick R., Blum S., Döring M., Giovanni R., Robertson T., Vieglais D. 2012. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7:e29715.

Wiens J.J., Collins T. 2004. The Role of Morphological Data in Phylogeny Reconstruction. Syst. Biol. 53:653–661.

Wiens J.J., Kuczynski C.A., Townsend T., Reeder T.W., Mulcahy D.G., Sites J.W. 2010. Combining Phylogenomics and Fossils in Higher-Level Squamate Reptile Phylogeny: Molecular Data Change the Placement of Fossil Taxa. Syst. Biol. 59:674–688.

Wipfler B., Pohl H., Yavorskaya M.I., Beutel R.G. 2016. A review of methods for analysing insect structures — the role of morphology in the age of phylogenomics. Curr. Opin. Insect Sci. 18:60–68.

Wood H.M., Matzke N.J., Gillespie R.G., Griswold C.E. 2013. Treating Fossils as Terminal Taxa in Divergence Time Estimation Reveals Ancient Vicariance Patterns in the Palpimanoid Spiders. Syst. Biol. 62:264–284.

Yassin A. 2013. Phylogenetic classification of the Drosophilidae Rondani (Diptera): the role of morphology in the postgenomic era. Syst. Entomol. 38:349–364.

Yesson C., Brewer P.W., Sutton T., Caithness N., Pahwa J.S., Burgess M., Gray W.A., White R.J., Jones A.C., Bisby F.A., Culham A. 2007. How Global Is the Global Biodiversity Information Facility? PLoS ONE 2:e1124.

Figure captions

Figure 1: Illustrations of observation-based and specimen-based primary biodiversity occurrences and their potential uses. a) Observations (top) and voucher specimens (bottom) can be complemented with ancillary data such as multimedia files or DNA sequences. For observations, these additional data must be acquired when the observation is performed; it cannot be performed later. On the opposite, for specimens – as long as they are well-curated, which is unfortunately not always possible – ancillary data can be gathered later (this feature is symbolized through the continuous background and the arrows). **b)** Three hypothetical case studies – Because data can be acquired later, a specimen occurrence offers a wide range of studies and analyses. Conversely, for observation occurrences, the spectrum of analyses depends on the existence or not of ancillary data: an unvouchered observation will not allow as many studies as an observation combined with a DNA sample (the interdiction signs cover studies that cannot be achieved). Pictograms for specimen, observation, DNA and photos were designed by FreepiK from Flaticon.

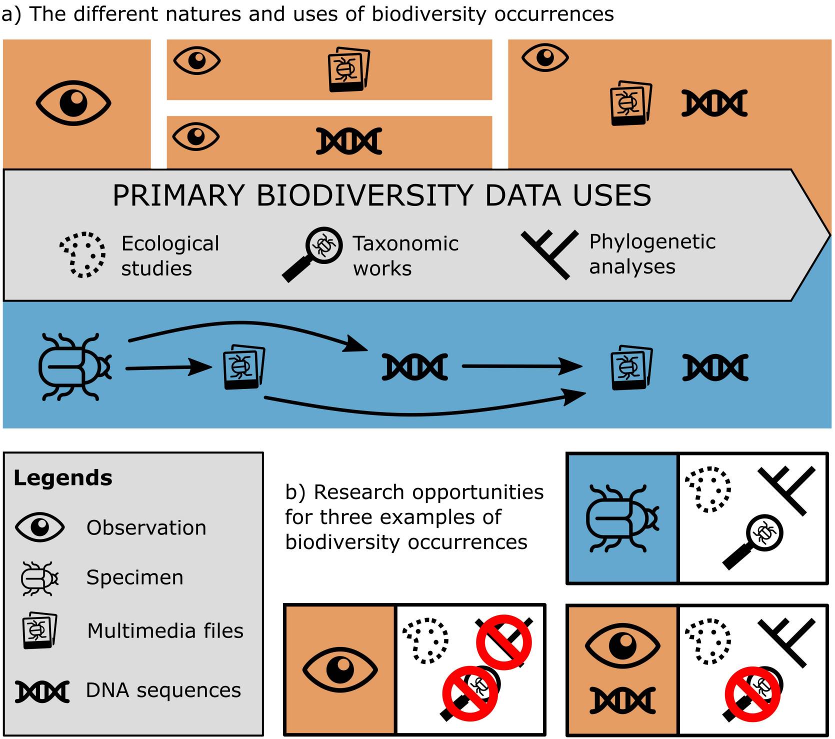
Figure 2: Number of primary biodiversity occurrences per year and origin from 1900 to today. The plot shows that observation-based occurrences have outnumbered specimen-based occurrences since 1970 and that this excess is growing. Occurrences from the last ten years are shaded because the pace at which data are added within the GBIF portal, especially for specimen-based occurrences, likely affects them.

Figure 3: Proportion of occurrences per year of collection and origin for a particular class. For each class, areas represent, from top to bottom, the proportions of specimen-based, observation-based and unknown origin occurrences. Contrary to 50 years ago, a majority of observation occurrences is reported whatever the taxonomic class.

Figure 4: The increase of ancillary data to biodiversity occurrences does not keep pace with biodiversity data accumulation. The top plot shows a yearly report of the number of multimedia files and DNA sequences linked to occurrences. The bottom plot shows the mean number of additional data per occurrence with multimedia files and DNA sequences.

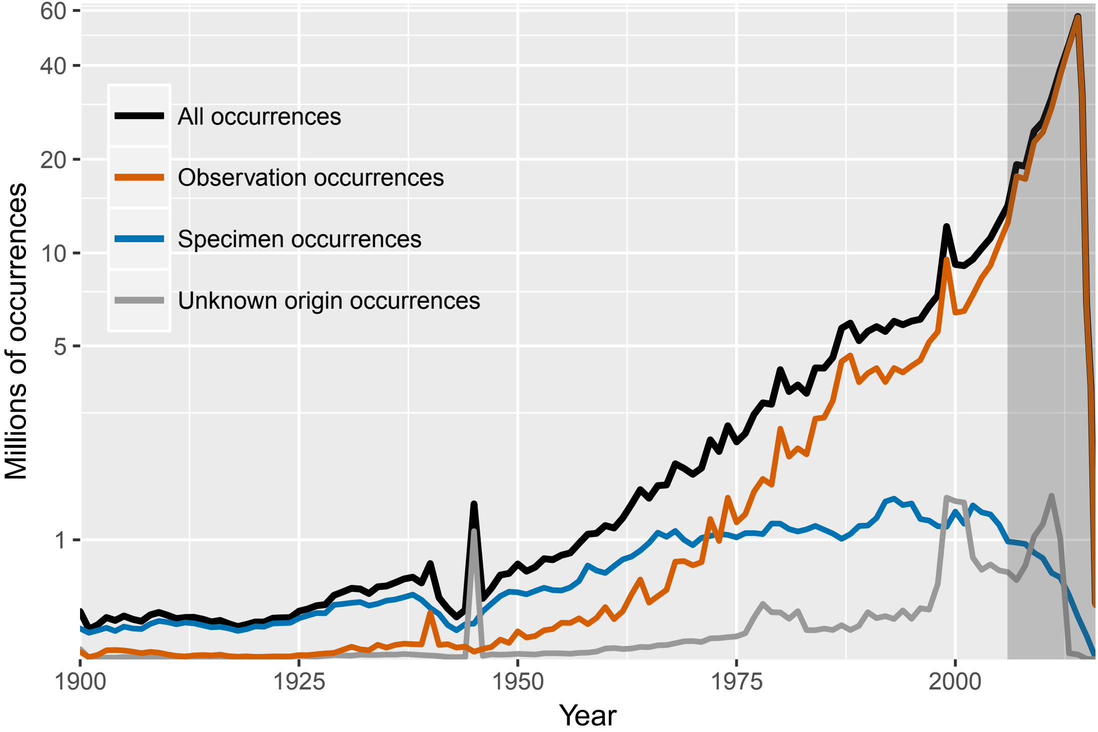
Figure 5: Occurrences with ancillary data are mainly specimen occurrences. Occurrences with multimedia files (*left*) are mainly specimen-based, whereas occurrences with DNA sequences (*right*) are either specimen-based or of unknown origin. Very few observations-based occurrences are provided with ancillary data.

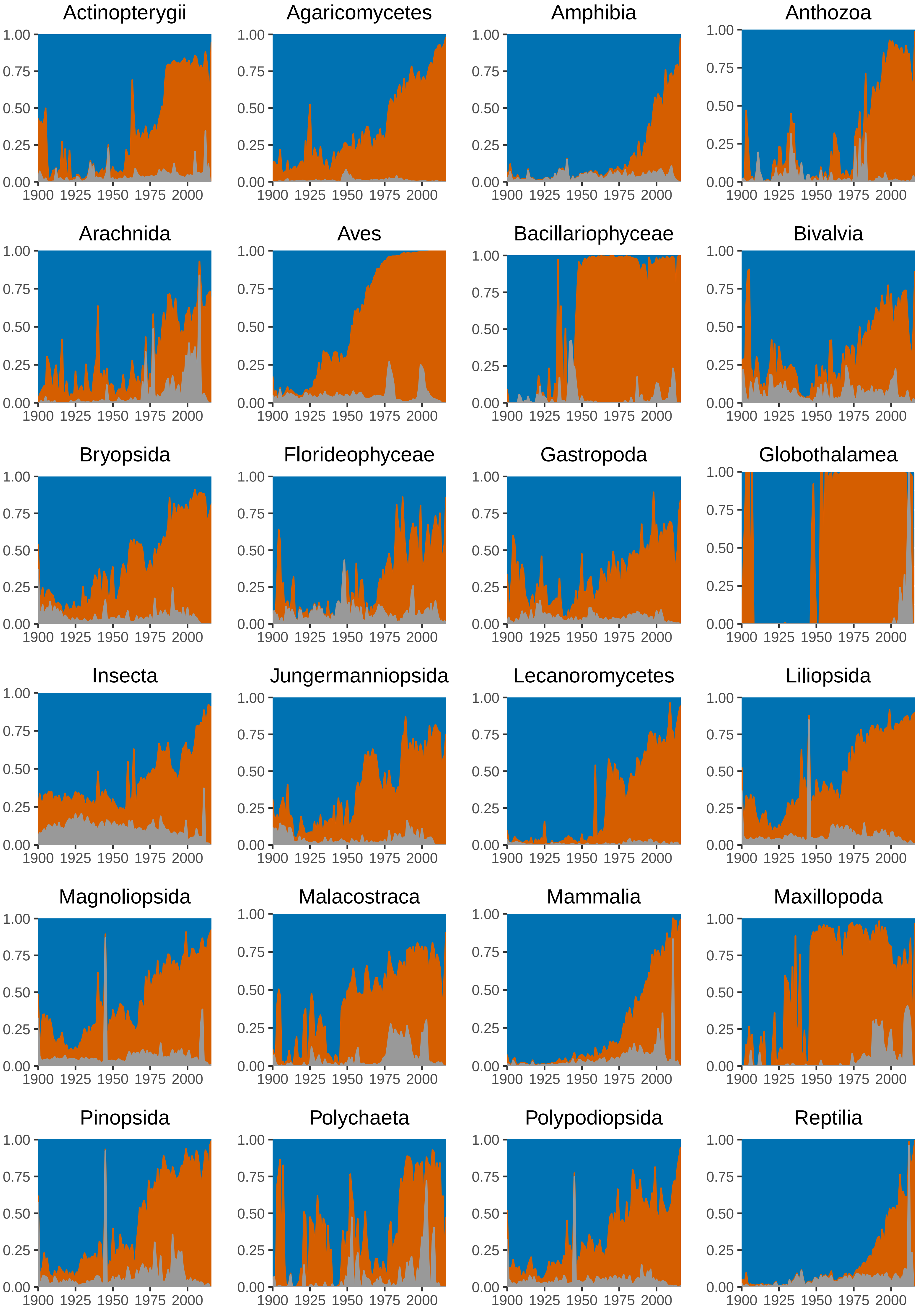
Figure 6: a) Spatial and b) taxonomic precision in the GBIF mediated data improve over time in proportion. The plot a) shows the number of occurrences collected each year lacking coordinates or tagged as having geospatial issues in the GBIF (plain line). Yet, the proportion of those occurrences is decreasing (dashed line). The plot b) shows the number of occurrences identified at least at the species level or at a higher taxonomic rank. The number of occurrences identified at a higher taxonomic rank is increasing with time. Yet, the proportion of occurrences identified at least at the species level is increasing.

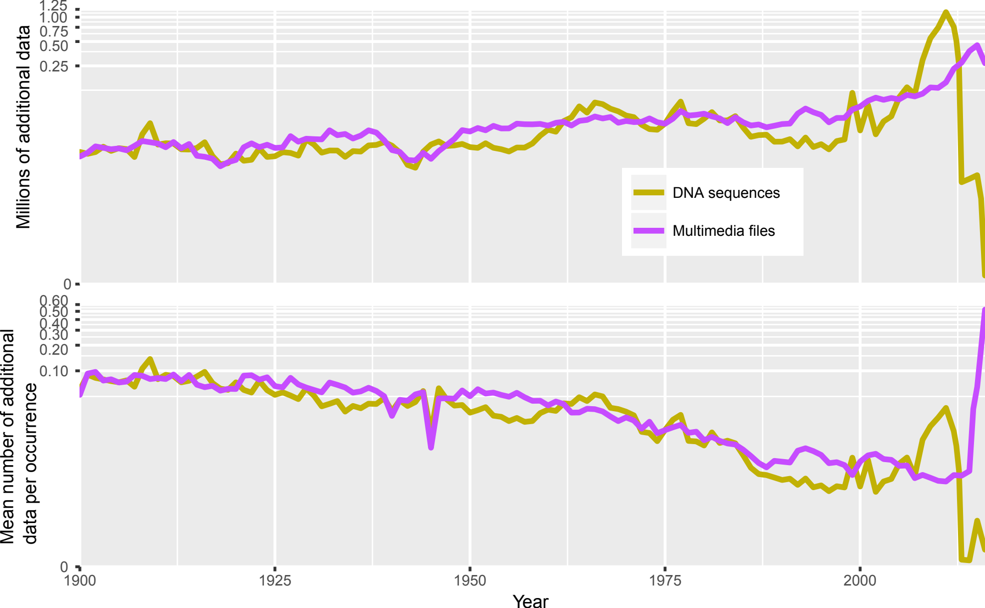


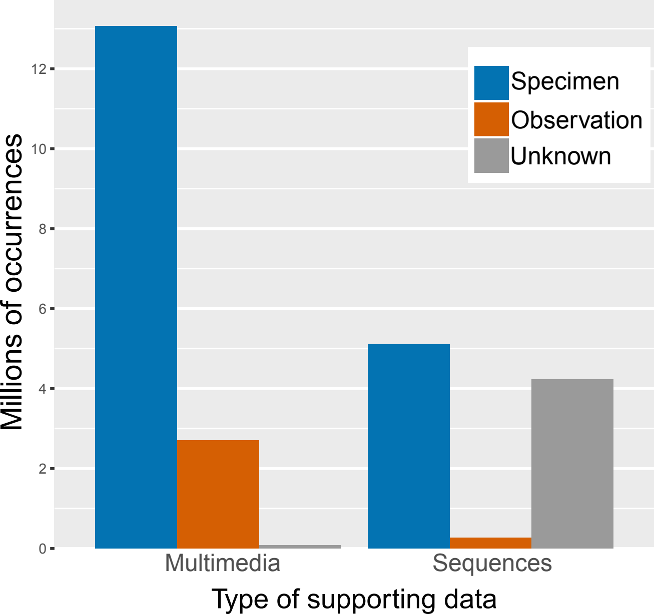
a) The different natures and uses of biodiversity occurrences

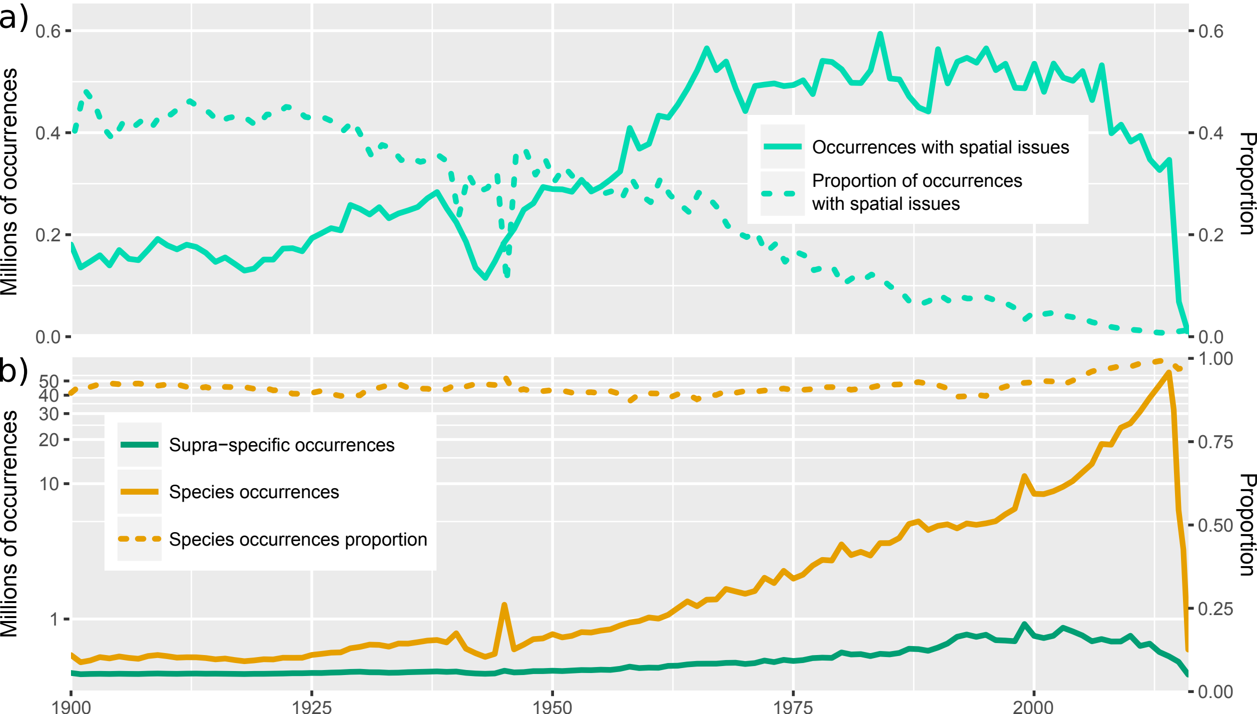
b) Research opportunities for three examples of biodiversity occurrences


All occurrences
Observation occurrences
Specimen occurrences
Unknown origin occurrences
Millions of occurrences
60
40
20
10
5
1
Year
1900
1925
1950
1975
2000


Actinopterygii
Agaricomycetes
Amphibia
Anthozoa
Arachnida
Aves
Bacillariophyceae
Bivalvia
Bryopsida
Florideophyceae
Gastropoda
Globothalamea
Insecta
Jungermanniopsida
Lecanoromycetes
Liliopsida
Magnoliopsida
Malacostraca
Mammalia
Maxillopoda
Pinopsida
Polychaeta
Polypodiopsida
Reptilia
0.00
0.25
0.50
0.75
1.00
1900
1925
1950
1975
2000


Millions of additional data
1.25
1.00
0.75
0.50
0.25
0
DNA sequences
Multimedia files
Mean number of additional data per occurrence
0.60
0.50
0.40
0.30
0.20
0.10
0
1900
1925
1950
1975
2000
Year


Specimen
Observation
Unknown
Millions of occurrences
0
2
4
6
8
10
12
Multimedia
Sequences
Type of supporting data


a)
Millions of occurrences
0.6
0.4
0.2
0.0
Proportion
Occurrences with spatial issues
Proportion of occurrences with spatial issues
b)
Millions of occurrences
50
40
30
20
10
1
Proportion
1.00
0.75
0.50
0.25
0.00
Supra−specific occurrences
Species occurrences
Species occurrences proportion
1900
1925
1950
1975
2000