



**HAL**  
open science

## 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

**HAL Id: hal-02314020**

**<https://hal.sorbonne-universite.fr/hal-02314020>**

Submitted on 11 Oct 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 ***Running head***

2 SHIFT IN PRIMARY BIODIVERSITY DATA

3

4 ***Title***

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

7 ***Authors' names***

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

9

10 ***Authors' affiliations***

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

13 *\*Correspondence to be sent to: Institut Systématique, Evolution, Biodiversité (ISYEB),*  
14 *Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, 57 rue Cuvier,*  
15 *75005 Paris, France; E-mail: [frederic.legendre@mnhn.fr](mailto:frederic.legendre@mnhn.fr)*

16

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

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

39

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

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

65

66           This change in practice (proportionally much more OB than SB occurrences) is a major  
67 upheaval with spectacular consequences for systematics and evolutionary studies. Since the very  
68 beginning of systematics, specimens have been collected and used to inventory the diversity of life  
69 and later to decipher the relationships within the tree of life (Giribet 2015). Natural history collections  
70 (NHC), which now support biodiversity, morphology or molecular databases, have been put together  
71 and used for species identification and description, comparative anatomy, and phylogenetic studies, to  
72 name a few practices embodying their usefulness (Garner et al. 2014; Kemp 2015; Buerki and Baker  
73 2016; Fig. 1). Databases containing mainly unvouchered observations would not be as profitable as  
74 data repositories composed of specimens but they have positive sides in return (e.g. the pace at which  
75 biodiversity occurrences are shared; datasets with higher statistical value, etc.) and can be  
76 complemented with diverse media. It is pivotal to accommodate to this shift now to make the best of  
77 it. As often, good legacy of previous practices and fruitful innovations must be retained and  
78 developed, while bad legacy must be put aside (Godfray 2002).

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

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

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

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

115

## 116 MATERIALS AND METHODS

### 117 *Data set*

118         We downloaded all the data available from the GBIF portal in June 2016  
119 (<http://doi.org/10.15468/dl.hqesx6>). These 649 million occurrences were saved as a Darwin Core

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

### 132 *Data Quantity*

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

### 135 *Data Origin*

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

### 145 *Supporting Files*

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

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

#### 157 *Development of Data Completeness*

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

#### 166 *Development of Taxonomic and Spatial Precision*

167            In general, a primary biodiversity occurrence is associated to a scientific name, which can be  
168 more or less precise depending on the skills of the identifier but also on the state and availability of  
169 taxonomic knowledge. We estimated taxonomic precision (in number and proportion per year)  
170 differentiating occurrences identified at least at the species level from supra-specific occurrences. The  
171 proportion of occurrences identified at the species or infraspecific level was used to estimate the



172 taxonomic precision of the GBIF mediated occurrences. As for the development of spatial  
173 imprecision, it was calculated as the number and proportion, per year, of occurrences lacking  
174 coordinates or flagged in the GBIF as data with coordinate issues.

## 175 RESULTS AND DISCUSSION

### 176 *A Shift in the Recording of Primary Biodiversity Data*

177 In the current context of biodiversity crisis, numerous pleas have incited the scientific  
178 community to collect as much biodiversity data as possible, out of the fear it might disappear before  
179 we even knew of its existence (May 2004; Butchart et al. 2010). These calls have been heard and,  
180 indisputably, biodiversity data accumulate faster than ever (Fig. 2 and Supporting Information), a  
181 trend most classes of organisms exhibit even though for a few of them the trend is not so strong  
182 (Troudet et al. 2017). The >57 million occurrences submitted to the GBIF in 2014, more than five  
183 times the amount of data submitted ten years earlier (i.e. 11 million occurrences in 2004), embody this  
184 report (Supporting Information). With this spectacular acceleration, the amount of data available to  
185 scientists is so huge that the study of biodiversity has entered into the “Big Data” era (Hampton et al.  
186 2013; Joppa et al. 2016; Kelling et al. 2009). Arguably, this trend reflects the advent of new scientific  
187 communities taking or renewing their interest in biodiversity. It may also suggest an increasing appeal  
188 of the public for biodiversity. In both cases, this situation offers new opportunities that will enrich our  
189 understanding of biodiversity and generate a higher awareness of conservation issues or of  
190 biodiversity shortfalls (Hortal et al. 2015). Other benefits followed such as an increased power in  
191 statistical analyses because of larger datasets or the possibility to tackle issues at large taxonomical,  
192 temporal or spatial scales (Rosenheim and Gratton 2017). However, the large volume of data is also a  
193 curation challenge that must be handled to avoid passing on a dubious source of knowledge to future  
194 generations because of a fall in data quality (Howe et al. 2008), a criticism regularly brought up for  
195 GBIF mediated data (e.g. Yesson et al. 2007).

196 This acceleration is triggered, at least partly, by a change in the way biodiversity data are  
197 recorded. The origin of biodiversity data has shifted from a majority of specimen-based (SB) to a

198 majority of observation-based (OB) occurrences. This shift has been previously suspected (Gaiji et al.  
199 2013) and we show here that, from 1970 to 2016, the proportion of occurrences traceable to tangible  
200 material (i.e. specimens) fell from 68 % to 18 %. This decrease is not due to recent digitization  
201 initiatives because we used the collection/observation date, and not the digitization date, to compute  
202 these statistics. This result applies to the 24 classes studied, except for a few eccentric cases such as  
203 Globothalamea and Polychaeta (Figs. 2 and 3). Likely, these exceptions relate to specific practices for  
204 observing, collecting or curating these organisms, or to their low volume of primary biodiversity data,  
205 which might cast doubt on their atypical trends. Besides, this shift might be slightly inflated because it  
206 presumably requires less time to integrate OB than SB occurrences in the GBIF. Still, ignoring the last  
207 ten years to limit this potential bias (shaded area in Fig. 2), this shift remains striking. It started, for  
208 most of the organisms, in the second half of the 20<sup>th</sup> century and kept intensifying ever since. On the  
209 opposite, the number of SB occurrences has stagnated, or increased marginally at best, in the past 40  
210 years. More worrying, most of SB occurrences cannot be readily traced back to a specimen: Only 238  
211 000 occurrences have a filled “*materialsamplid*” column, representing only 0.28 % of the 84 million  
212 SB occurrences. This number illustrates that the way SB occurrences are recorded in biodiversity  
213 databases must be improved. Even though a specimen exists somewhere, it cannot be located without  
214 a potentially complex investigation procedure. This situation hampers the verification process, a  
215 founding step in scientific practice (Turney et al. 2015). Although scientists can be delighted with the  
216 pace at which biodiversity data accumulate, they cannot be satisfied with a biodiversity research  
217 relying mainly on unverifiable or hardly verifiable occurrences.

218         Divergent causes, not necessarily exclusive, may explain this practice shift. In a context of  
219 massive biodiversity loss, a sense of urgency fueled the pleas for accelerated data collection (Hampton  
220 et al. 2013) and encouraged the accumulation of unvouchered observations, less destructive and easier  
221 to produce, share and store than specimen-based occurrences. Ethical considerations and conservation  
222 issues that hinder specimen collections have commonly been put forward (e.g. Minter et al. 2014),  
223 although they are debatable in some situations (Dubois and Nemésio 2007; Dubois 2009; Dubois  
224 2017; Löbl 2017). The adoption of the Nagoya protocol by many countries and the development of

225 mobile applications have undoubtedly contributed to this shift as well. Concurrently, Grandcolas  
226 (2017) suggested that this shift started when biodiversity sciences merged with general biology, more  
227 interested in discovering general patterns and laws than in documenting diversity that is supposedly  
228 already known enough. Others underlined the lack of human and economic resources to ensure both  
229 the gathering of specimens and the curation of natural history collections (Kemp 2015). These reasons  
230 could have favoured a decrease in specimen collection. On the other hand, the number of observation-  
231 based occurrences has dramatically increased with, for instance, the rise of citizen science that enable  
232 to rapidly produce a vast amount of observational data (Dickinson et al. 2012) and that will certainly  
233 become more attracting and rewarding for the public in the future (e.g. Silvertown et al. 2015). Given  
234 the multiple origins of this trend, it seems unlikely to be reversed in the near future and must be  
235 organised and guided to ensure maximal benefits for the study of biodiversity.

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

238 The importance of collecting specimens in taxonomy, evolution and ecology cannot be  
239 overemphasized (Huber 1998; Schilthuizen et al. 2015) and two main points, previously discussed in  
240 the literature, must be reiterated. First, specimens are needed for species description and for the study  
241 of biodiversity in general (Krell and Wheeler 2014; Rocha et al. 2014; Ceriaco et al. 2016; Dubois  
242 2017; Gutiérrez and Pine 2017; Pine and Gutiérrez 2018 *contra* Minter et al. 2014; Marshall and  
243 Evenhuis 2015; Pape et al. 2016). A crucial argument is the utility of specimens for checking species  
244 identification. The spectacular growth in biodiversity occurrences is a fantastic opportunity for  
245 researchers if, and only if, occurrence quality can be somehow evaluated. Goodwin et al. (2015)  
246 assessed that up to half of tropical plant identifications in museum collections were false. Correcting  
247 identification errors can be done after examining specimens, but is impossible for unvouchered  
248 observations. If Goodwin et al.'s estimation is correct and generalizable to most primary data, the need  
249 for some specimens, or at least ancillary data to observation occurrences, is critical. Encouragingly,  
250 Kosmala et al (2016) showed that high quality data were obtained in diverse citizen science programs,  
251 suggesting that biodiversity data gathered for well-known taxa (Troudet et al. 2017) or geographic

252 areas (Meyer et al. 2016) might contain fewer errors than in the study of Goodwin et al. (2015).  
253 Second, the revived focus on morphology advocated lately in systematics requires specimens (Jenner  
254 2004, Wiens 2004, Smith and Turner 2005, Yassin 2013, Pyron 2015, Wanninger 2015, Wipfler et al.  
255 2016). Authors recommending this revival underlined that comparative morphology not only brings  
256 phylogenetic characters but also allows including fossil taxa in phylogenetic analyses (e.g. Pyron  
257 2011; Wood et al. 2013), enabling us to better estimate the structure and branch length of the  
258 reconstructed trees (Wiens *et al.* 2010; Pyron 2015). Given that phylogenetic thinking has become of  
259 paramount importance in biology, improvements in phylogenetic estimation offer large potentialities  
260 in comparative analyses and evolutionary studies, and in the study of biodiversity in general (Losos et  
261 al. 2013; Buerki et al. 2015).

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

269 A specimen is not always necessary for a primary biodiversity data to be useful. Instead of  
270 specimens, and in complement to unvouchered observations, digital data or molecular data can be  
271 collated. New technologies offer a wide range of tools and methods to collect concrete specimen  
272 evidence in nature, and it is now relatively easy and affordable to obtain DNA sequences, images and  
273 sound recordings. Then, using molecular and digital data should now be a common practice in the  
274 study of biodiversity, as the exponential growth of molecular data and phylogenies, and the  
275 development of morphological databases and ontogenies would suggest (Lathe et al. 2008; Parr et al.  
276 2012; Deans et al. 2012, 2015). We show here that digital and DNA data are increasingly used (63,271  
277 and 878,308 ancillary data were collected in 1950 and 2010, respectively) but these data remain  
278 patently underemployed (Fig. 4). Only 2.5 % of all the GBIF-mediated occurrences for the 24 focal

279 classes were linked to digital data and 1.5 % to DNA sequences. Worse, proportionally, they become  
280 more and more negligible, regarding the large quantity of observations without supporting data. This  
281 situation might be improving lately, but the post-2008 tendency observed demands to be confirmed in  
282 future years (Fig. 4), which would be happening only if scholars take up the pedagogic and practical  
283 challenge of highlighting the importance of ancillary data for biodiversity occurrences. Moreover, and  
284 quite inconsistently, digital and DNA data were less used for OB than for SB occurrences (Fig. 5).  
285 They would yet be more useful for OB biodiversity data given that they would constitute the only way  
286 to independently check or update observation occurrences, whereas one can refer to specimens, as  
287 long as those are kept and the traceability chain is not broken, for SB occurrences (Page 2015; Nualart  
288 et al. 2017). The high proportion of sequences associated to primary biodiversity data of unknown  
289 origin could suggest that when a sample is performed, occurrences are often classified in the catch-all  
290 class 'unknown origin'.

291 In addition to ancillary data, the usefulness of primary biodiversity occurrences can be  
292 maximized through a higher level of precision and completeness in recordings. We expect biodiversity  
293 data occurrences to be more precise and complete now than before because tools that are more  
294 efficient have been developed. Whatever the nature of the occurrence, spatial coordinates for instance  
295 can be easily provided with a high precision level given the democratization of GPS. Data  
296 completeness should also improve because of the growing awareness that a global and comprehensive  
297 picture of biodiversity is needed. Our results showed that, in proportion, data precision does improve  
298 but that data completeness stagnates (Fig. 6 and Supporting Information). The proportion of data with  
299 geospatial issues in the GBIF (i.e. data with low spatial precision) decreased from 50.2 % in 1900 to  
300 0.6 % in 2014 in spite of a larger number of occurrences with spatial imprecision – this number being  
301 quite stable over the past 30 years (Fig. 6A). Over the same period, records identified at the species  
302 level augmented from 89.6% to 99.4%, with once again an increase of supra-species records (Fig. 6B).  
303 While species identification and spatial precision improves, so does niche modelling results for  
304 instance, which promises significant advances in biogeography (e.g. Meseguer et al. 2015; Töpel et al.

305 2017). In this regard, important gains for systematics and evolutionary studies can be anticipated from  
306 the increasing level of precision in primary biodiversity data.

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

324         The fear of biodiversity disappearance has triggered a vague of biodiversity data accumulation.  
325 We are in the middle of a paradigm change where most biodiversity data are not anymore gathered  
326 like it used to be. This paradigm change has been undergone without any supervision. Even though  
327 some aspects of these changes are highly beneficial, others are suboptimal and must not be ignored.  
328 We must act now to allow a better monitoring of the biodiversity research agenda and to continue  
329 shaping how biodiversity data should be gathered, diversifying the objects of collection (e.g.  
330 specimens, samples, DNA, images, etc. – Knapp 2015). We argue that ancillary data (samples, DNA,  
331 pictures) must be collected more methodically than today (Joppa et al. 2016), to avoid disillusionment

332 when we will realize that unvouchered observations were not sufficient to address some current and  
333 future preoccupying issues about systematics and evolutionary studies.

334

#### 335 ACKNOWLEDGMENTS

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

343

#### 344 REFERENCES

- 345 Anmarkrud J.A., Lifjeld J.T. 2017. Complete mitochondrial genomes of eleven extinct or  
346 possibly extinct bird species. *Mol. Ecol. Resour.* 17:334–341.
- 347 Ariño A.H. 2010. Approaches to estimating the universe of natural history collections data.  
348 *Biodiv. Inf.* 7:81–92
- 349 Bisby F.A. 2000. The Quiet Revolution: Biodiversity Informatics and the Internet. *Science*  
350 289:2309–2312.
- 351 Buerki S., Baker W.J. 2016. Collections-based research in the genomic era. *Biol. J. Linn. Soc.*  
352 117:5–10.

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

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

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

366 Ceriáco, L.M.P., Gutiérrez, E.E., Dubois, A. et al. 2016. Photography-Based Taxonomy Is  
367 Inadequate, Unnecessary, and Potentially Harmful for Biological Sciences. *Zootaxa*  
368 4196:435–45.

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



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

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

383 Dickinson J.L., Shirk J., Bonter D., Bonney R., Crain R.L., Martin J., Phillips T., Purcell K.  
384 2012. The current state of citizen science as a tool for ecological research and public  
385 engagement. *Front. Ecol. Environ.* 10:291–297.

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

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

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

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

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

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

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

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

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

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

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

407 Gutiérrez, E.E., Pine, R.H. 2017. Specimen collection crucial to taxonomy. *Science* 6331:  
408 1275–1275.

409 Hampton S.E., Strasser C.A., Tewksbury J.J., Gram W.K., Budden A.E., Batcheller A.L.,  
410 Duke C.S., Porter J.H. 2013. Big data and the future of ecology. *Front. Ecol. Environ.*  
411 11:156–162.

412 Hortal J., Bello F. de, Diniz-Filho J.A.F., Lewinsohn T.M., Lobo J.M., Ladle R.J. 2015. Seven  
413 shortfalls that beset large-scale knowledge of biodiversity. *Annu. Rev. Ecol. Evol. Syst.*  
414 46:523–549.

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

418 Huber J.T. 1998. The importance of voucher specimens, with practical guidelines for  
419 preserving specimens of the major invertebrate phyla for identification. *J. Nat. Hist.*  
420 32:367–385.

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

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

426 Kelling S., Hochachka W.M., Fink D., Riedewald M., Caruana R., Ballard G., Hooker G.  
427 2009. Data-Intensive Science: A New Paradigm for Biodiversity Studies. *BioScience*  
428 59:613–620.

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

430 Kitchin R. 2014. Big Data, new epistemologies and paradigm shifts. *Big Data & Society.*  
431 1:2053951714528481.

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

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

437 Krell F.T., Wheeler Q.D. 2014. Specimen collection: Plan for the future. *Science* 344:815–  
438 816.

- 439 Lathe W., Williams J., Mangan M., Karolchik D. 2008. Genomic data resources: challenges  
440 and promises. *Nat. Educ.* 1:2.
- 441 Löbl I. 2017. Assessing biodiversity: a pain in the neck. *Bionomina* 12:39–43.
- 442 Losos J.B., Arnold S.J., Bejerano G., Iii E.D.B., Hibbett D., Hoekstra H.E., Mindell D.P.,  
443 Monteiro A., Moritz C., Orr H.A., Petrov D.A., Renner S.S., Ricklefs R.E., Soltis P.S.,  
444 Turner T.L. 2013. Evolutionary Biology for the 21st Century. *PLoS Biol.* 11:e1001466.
- 445 Marshall S.A., Evenhuis N.L. 2015. New species without dead bodies: a case for photo-based  
446 descriptions, illustrated by a striking new species of *Marleyimyia* Hesse (Diptera,  
447 Bombyliidae) from South Africa. *Zookeys* 525:117–127.
- 448 May R.M. 1990. Taxonomy as destiny. *Nature* 347:129–130.
- 449 May R.M. 2004. Tomorrow’s taxonomy: collecting new species in the field will remain the  
450 rate-limiting step. *Philos. T. Roy. Soc. B.* 359:733–734.
- 451 Marx V. 2013. Biology: The big challenges of big data. *Nature* 498:255–260.
- 452 Meseguer A.S., Lobo J.M., Ree R., Beerling D.J., Sanmartín I. 2015. Integrating Fossils,  
453 Phylogenies, and Niche Models into Biogeography to Reveal Ancient Evolutionary  
454 History: The Case of *Hypericum* (Hypericaceae). *Syst. Biol.* 64:215–232.
- 455 Meyer C., Weigelt P., Kreft H., Lambers J.H.R. 2016. Multidimensional biases, gaps and  
456 uncertainties in global plant occurrence information. *Ecol. Lett.* 19:992–1006.
- 457 Minter B.A., Collins J.P., Love K.E., Puschendorf R. 2014. Avoiding (Re)extinction. *Science*  
458 344:260–261.
- 459 Nualart N., Ibáñez N., Soriano I., López-Pujol J. 2017. Assessing the Relevance of Herbarium  
460 Collections as Tools for Conservation Biology. *Bot. Rev.* 83:303–325.

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

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

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

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

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

473 Pine, R.H., Gutiérrez, E.E. 2018. What Is an ‘Extant’ Type Specimen? Problems Arising from  
474 Naming Mammalian Species-Group Taxa without Preserved Types. Mammal Review  
475 48:12–23.

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

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

480 Rocha L.A., Aleixo A., Allen G., Almeda F., Baldwin C.C., Barclay M.V.L., Bates J.M.,  
481 Bauer A.M., Benzoni F., Berns C.M., Berumen M.L., Blackburn D.C., Blum S., Bolaños  
482 F., Bowie R.C.K., Britz R., Brown R.M., Cadena C.D., Carpenter K., Ceriaco L.M.,

483 Chakrabarty P., Chaves G., Choat J.H., Clements K.D., Collette B.B., Collins A., Coyne  
484 J., Cracraft J., Daniel T., de Carvalho M.R., de Queiroz K., Di Dario F., Drewes R.,  
485 Dumbacher J.P., Engilis A., Erdmann M. V., Eschmeyer W., Feldman C.R., Fisher B.L.,  
486 Fjeldså J., Fritsch P.W., Fuchs J., Getahun A., Gill A., Gomon M., Gosliner T., Graves  
487 G.R., Griswold C.E., Guralnick R., Hartel K., Helgen K.M., Ho H., Iskandar D.T.,  
488 Iwamoto T., Jaafar Z., James H.F., Johnson D., Kavanaugh D., Knowlton N., Lacey E.,  
489 Larson H.K., Last P., Leis J.M., Lessios H., Liebherr J., Lowman M., Mahler D.L.,  
490 Mamonekene V., Matsuura K., Mayer G.C., Mays H., McCosker J., McDiarmid R.W.,  
491 McGuire J., Miller M.J., Mooi R., Mooi R.D., Moritz C., Myers P., Nachman M.W.,  
492 Nussbaum R.A., Foighil D.Ó., Parenti L.R., Parham J.F., Paul E., Paulay G., Pérez-Emán  
493 J., Pérez-Matus A., Poe S., Pogonoski J., Rabosky D.L., Randall J.E., Reimer J.D.,  
494 Robertson D.R., Rödel M.-O., Rodrigues M.T., Roopnarine P., Rüber L., Ryan M.J.,  
495 Sheldon F., Shinohara G., Short A., Simison W.B., Smith-Vaniz W.F., Springer V.G.,  
496 Stiasny M., Tello J.G., Thompson C.W., Trnski T., Tucker P., Valqui T., Vecchione M.,  
497 Verheyen E., Wainwright P.C., Wheeler T.A., White W.T., Will K., Williams J.T.,  
498 Williams G., Wilson E.O., Winker K., Winterbottom R., Witt C.C. 2014. Specimen  
499 collection: An essential tool. *Science* 344:814 LP-815.

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

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

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

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

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

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

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

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

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

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

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

- 527 Wipfler B., Pohl H., Yavorskaya M.I., Beutel R.G. 2016. A review of methods for analysing  
528 insect structures — the role of morphology in the age of phylogenomics. *Curr. Opin.*  
529 *Insect Sci.* 18:60–68.
- 530 Wood H.M., Matzke N.J., Gillespie R.G., Griswold C.E. 2013. Treating Fossils as Terminal  
531 Taxa in Divergence Time Estimation Reveals Ancient Vicariance Patterns in the  
532 Palpimanoid Spiders. *Syst. Biol.* 62:264–284.
- 533 Yassin A. 2013. Phylogenetic classification of the Drosophilidae Rondani (Diptera): the role  
534 of morphology in the postgenomic era. *Syst. Entomol.* 38:349–364.
- 535 Yesson C., Brewer P.W., Sutton T., Caithness N., Pahwa J.S., Burgess M., Gray W.A., White  
536 R.J., Jones A.C., Bisby F.A., Culham A. 2007. How Global Is the Global Biodiversity  
537 Information Facility? *PLoS ONE* 2:e1124.



538 **Figure captions**

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

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

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

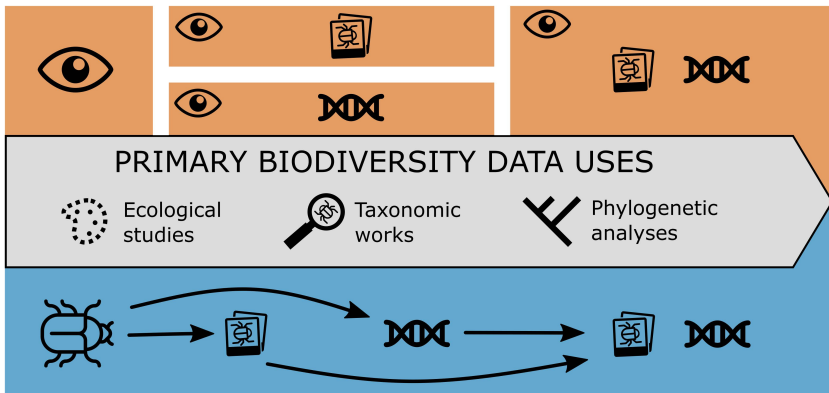
560 **Figure 4: The increase of ancillary data to biodiversity occurrences does not keep pace with**  
561 **biodiversity data accumulation.** The top plot shows a yearly report of the number of multimedia files  
562 and DNA sequences linked to occurrences. The bottom plot shows the mean number of additional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

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

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

575

a) The different natures and uses of biodiversity occurrences



**Legends**



Observation



Specimen

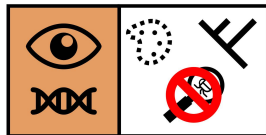
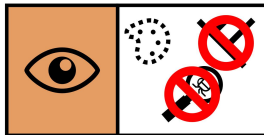
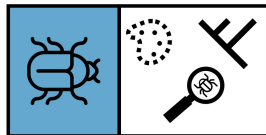


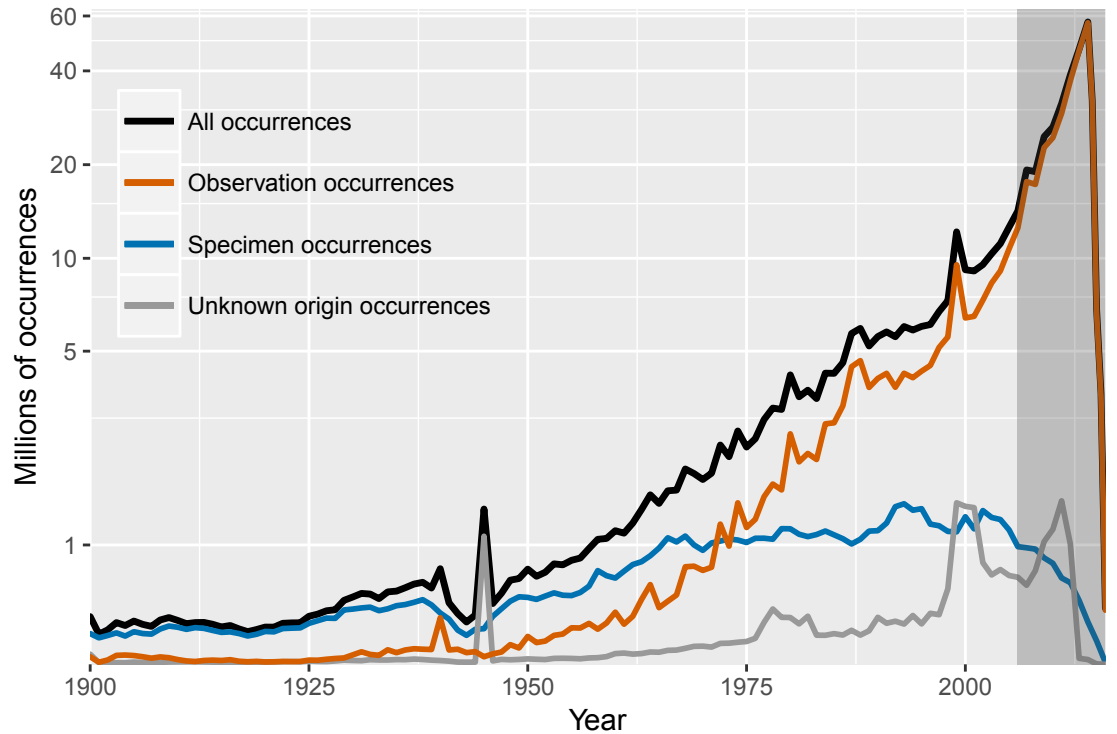
Multimedia files

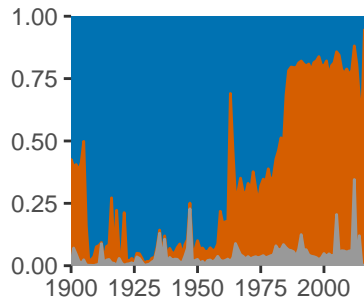
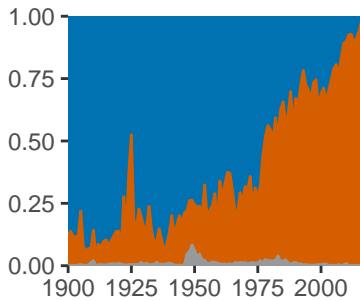
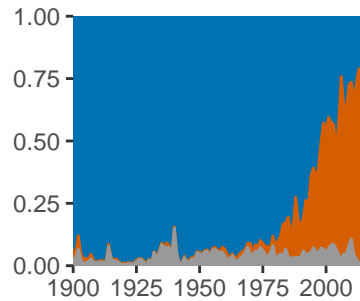
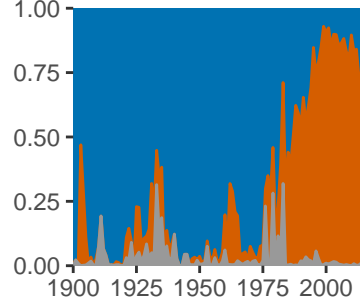
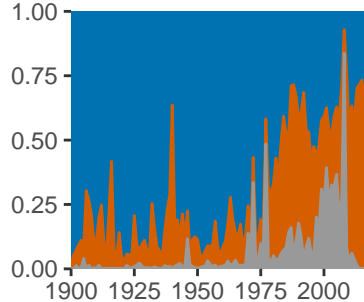
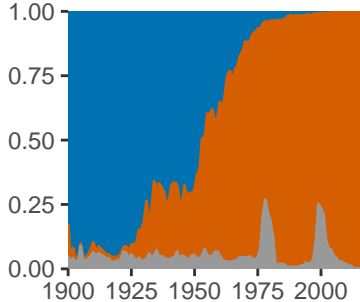
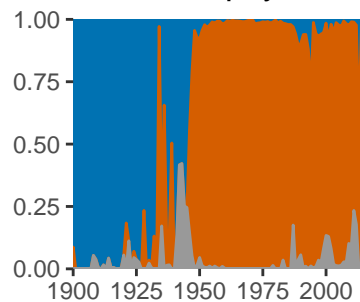
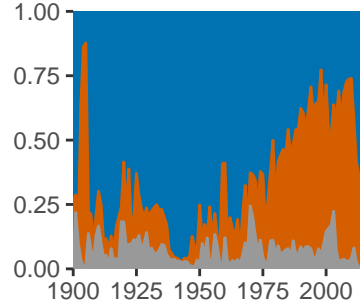
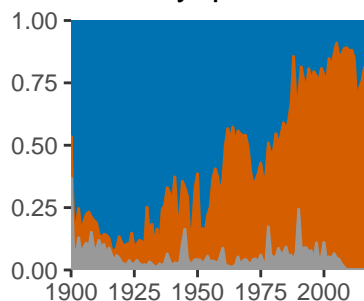
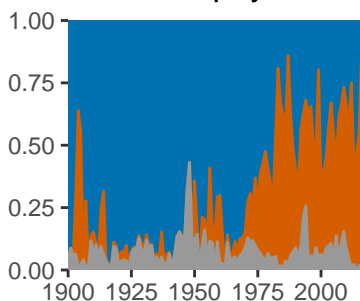
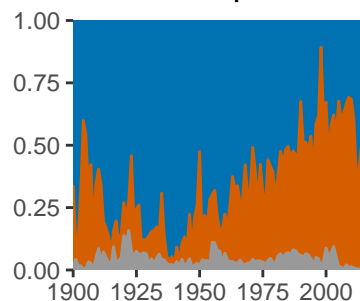
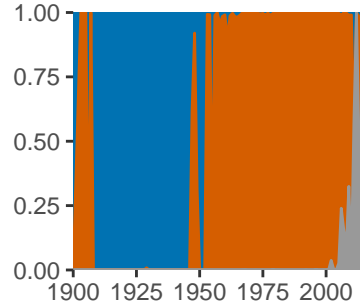
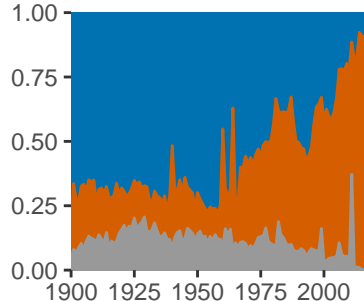
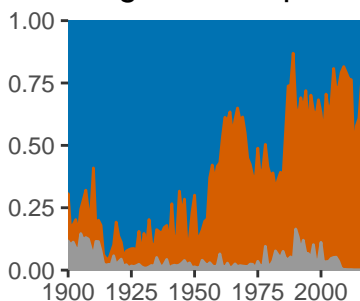
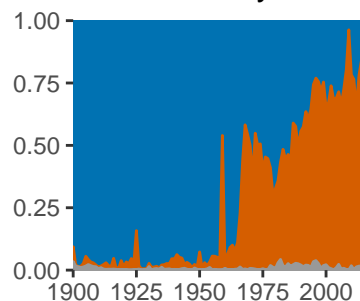
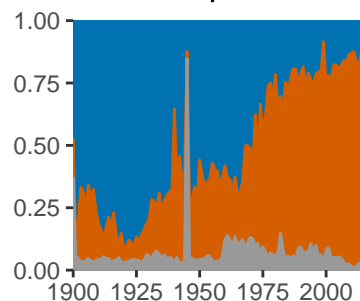
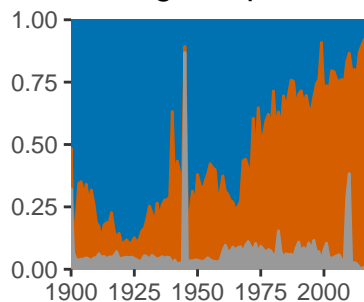
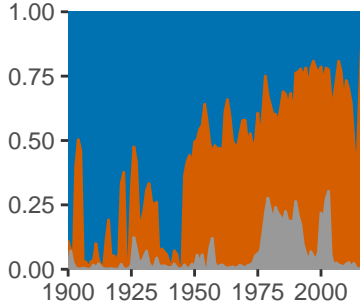
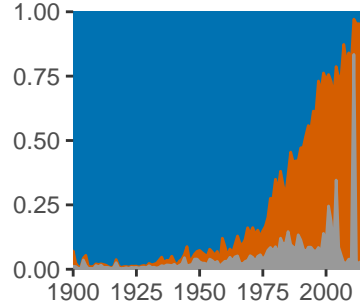
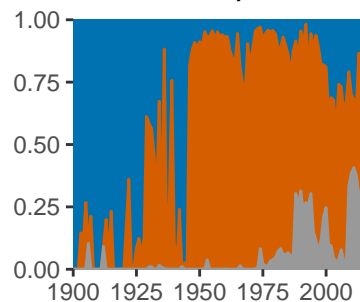
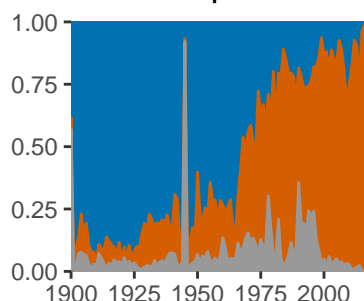
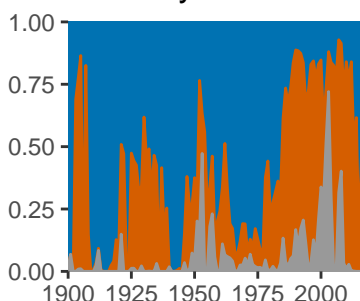
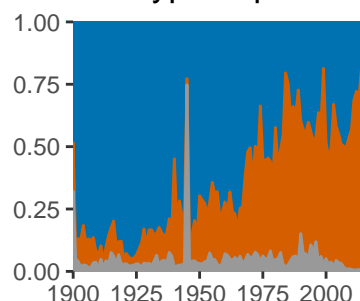


DNA sequences

b) Research opportunities for three examples of biodiversity occurrences





**Actinopterygii****Agaricomycetes****Amphibia****Anthozoa****Arachnida****Aves****Bacillariophyceae****Bivalvia****Bryopsida****Florideophyceae****Gastropoda****Globothalamea****Insecta****Jungermanniopsida****Lecanoromycetes****Liliopsida****Magnoliopsida****Malacostraca****Mammalia****Maxillopoda****Pinopsida****Polychaeta****Polypodiopsida****Reptilia**