

## HOW LONG DOES IT TAKE TO DISCOVER A SPECIES?

Journal:	<i>Systematics and Biodiversity</i>
Manuscript ID	TSAB-2020-0024.R1
Manuscript Type:	Original Research Article
Keywords:	Global plant inventory, herbaria, species discovery, specimen, taxonomy, tropical biodiversity

SCHOLARONE™  
Manuscripts

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**How long does it take to discover a species?**

ZOË A. GOODWIN<sup>1¶</sup>, PABLO MUÑOZ-RODRÍGUEZ<sup>2¶</sup>, DAVID J. HARRIS<sup>1</sup>, TOM  
WELLS<sup>2</sup>, JOHN R.I. WOOD<sup>2,3</sup>, DENIS FILER<sup>2</sup> AND ROBERT W. SCOTLAND<sup>2\*</sup>.

<sup>1</sup> Royal Botanic Garden Edinburgh, 20A Inverleith Row, EH3 5LR Edinburgh, United  
Kingdom

<sup>2</sup> Department of Plant Sciences, University of Oxford, South Parks Road, OX1 3RB Oxford,  
United Kingdom

<sup>3</sup> Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AB, United Kingdom.

\*Corresponding author  
Email: [robert.scotland@plants.ox.ac.uk](mailto:robert.scotland@plants.ox.ac.uk) (RS)

¶ These authors contributed equally to this work.

## Abstract

The description of a new species is a key step in cataloguing the World's flora. However, this is only a preliminary stage in a long process of understanding what that species represents. We investigated how long the species discovery process takes by focussing on three key stages: 1, the collection of the first specimen; 2, the publication of the species name; and 3, the date when a minimum number of fifteen accurately named specimens are available. We quantified the time lags associated with these stages for several groups of plants with different number of species and from different regions, including the 20 most species-rich Angiosperm families. Our analyses reveal that it takes decades to accumulate a minimum number of specimens to allow subsequent studies of any kind. The time lag between stages 1 and 3 is consistently over 70 years, with groups such as the tropical genus *Aframomum* reflecting an average time lag of over 100 years. In light of our results, we suggest that species discovery is most accurately characterized as a lengthy process of knowledge accumulation, often spanning decades, rather than a one-off event.

**Key words:** Global plant inventory, herbaria, species discovery, specimen, taxonomy, tropical biodiversity.

## Introduction

Knowledge of species and their characteristics is intertwined with almost every aspect of human activity. Maintenance of planetary health (Steffen et al., 2015), human welfare (Kraft et al., 2017), food security (Castañeda-Álvarez et al., 2016) and biodiversity conservation (Mace, 2004) all depend to some degree on our understanding of the characteristics of individual species. However, our knowledge of many plant species, especially in the tropics, is at best provisional, curtailing our capacity to achieve policy initiatives emanating from the

1  
2  
3 41 Convention on Biological Diversity (CBD, 2010). Even for a supposedly well-known group of  
4  
5 42 organisms such as flowering plants, it has been estimated that more than half of all specimens  
6  
7 43 in the world's herbaria do not have an accurate name (Goodwin et al., 2015), reflecting the  
8  
9 44 limited knowledge we have (Enquist et al., 2019; Nic Lughadha et al., 2019; Raven, 2007).

10  
11  
12 45 The global inventory of species is often considered to have two components: species that  
13  
14 46 have already been named and those that remain to be named (Costello et al., 2013). However,  
15  
16 47 it is becoming clear that the discovery of animal and plant species is a dynamic process drawn  
17  
18 48 out over time with, on average, 21 - 39 years between collecting the first specimen to its  
19  
20 49 publication as a new species (Bebber et al., 2010; Fontaine et al., 2012). In the case of flowering  
21  
22 50 plants, it has been shown that only 14% of new species are published within 5 years of the first  
23  
24 51 specimen being collected (Bebber et al., 2010). In addition, at the time of publication, many  
25  
26 52 new species are based on only a handful of specimens that offer minimal knowledge of their  
27  
28 53 distribution, frequency and ecology. In other words, species knowledge is probably best  
29  
30 54 thought of as a continuum from the unknown to well-known with many intermediate stages.

31  
32  
33 55 In this paper, we seek to quantify and measure the time lags associated with different  
34  
35 56 stages of the species discovery process. For this exercise, we recognise three key stages in the  
36  
37 57 early process of accumulating knowledge (Fig. 1): 1, the collection of the first specimen; 2, the  
38  
39 58 publication of the species name; and 3, the date when fifteen accurately named specimens are  
40  
41 59 available. Our choice of 15 accurately determined specimens to represent the third stage is  
42  
43 60 somewhat arbitrary but guided by consideration that this is a minimum level of knowledge  
44  
45 61 necessary for any reasonable level of understanding of the geographical distribution of a  
46  
47 62 species, with the exception of very restricted endemic species. The two time-intervals  
48  
49 63 associated with the three stages of species discovery that we recognise are referred to as *time*  
50  
51 64 *lag 1* (from first specimen collected to publication of the new species) and *time lag 2* (from  
52  
53 65 first specimen collected to 15 correctly determined specimens) (Fig. 1).

## 66 **Materials and methods**

### 67 **Specimens cited at the time of publication**

68 To explore how the number of specimens of a species accumulate over time, we first  
69 used the journal *Kew Bulletin* to calculate the number of specimens cited in the protologue  
70 when new species are published. We chose *Kew Bulletin* due to the large number of species  
71 described therein and because a lengthy period of publication was available. We studied all  
72 issues of the journal from 40 years (1970 - 2010), from which we extracted the number of  
73 specimens cited in every protologue of all newly published species of seed plants (Dataset 1).

74 To answer the question ‘how many specimens are cited when a species name is  
75 published?’, we calculated the mean number and standard deviation of specimens cited per  
76 species in the *Kew Bulletin* data from 1970 to 2010.

### 77 **Time lags of species discovery**

#### 78 **Time lags in *Aframomum***

79 We subsequently calculated the time lags described in the introduction for different  
80 groups of plants. We first calculated these time lags for the species in the genus *Aframomum*  
81 K.Schum., a tropical genus of herbaceous plants recently monographed (Harris & Wortley,  
82 2018). We used all species protologues together with label information from all specimens  
83 studied by Harris and Wortley between 1998 and 2015 at the Royal Botanic Garden Edinburgh.  
84 This dataset (Dataset 2) includes all *Aframomum* species protologues and specimen label data  
85 and determination slip information from all 3,176 herbarium collections (4,550 including  
86 duplicates) from 40 herbaria in 21 countries.

87 We classified the names of all determinations as either ‘correct’, ‘synonym’,  
88 ‘indeterminate’ or ‘other’ relative to the determination in the monograph (Harris & Wortley,  
89 2018). We treated basionyms of the correct name based on *Aframomum* as the correct name.  
90 We filtered the specimen and determination data to include only specimens for which both the

year of collection and the year of every determination are recorded (1,492 collections, 1,779 specimen duplicates). We excluded *A. angustifolium* (Sonn.) K.Schum. and *A. citratum* (J.Pereira) K.Schum. from all analyses because the date of collection for the first specimen of these two species is not accurately known. For every specimen, we recorded how long it took to collect 15 specimens (including misidentifications) and how long it took to accumulate 15 *correctly* determined specimens.

Also, when counting the time lags associated with each species from first collection to the various subsequent stages, we wanted to investigate whether there were different patterns for widespread, medium and restricted species *sensu* (Hawthorne & Marshall, 2016) as well as species with many specimens compared to species with few available specimens. We therefore calculated the mean of each time lag for geographically widespread, medium or restricted *Aframomum* species which occupy >24, 24 - 9 and 9 degree squares respectively.

In addition, when counting the period of time between the collection of the first and the 15<sup>th</sup> specimen of a species, we considered in our calculations the fact that, when a specimen is collected, there is usually a delay until the specimen is correctly identified and determined. To document this delay, we calculated time lag 2 in two complementary ways for each species recognised by Harris & Wortley (2018). First, we calculated the time lag between the first collection and the date of collection of the 15<sup>th</sup> specimen of a given species, irrespective of whether the specimens had the correct name at the time of collection, which was seldom the case. Secondly, we calculated the time lag between the first collection and the *correct* determination of the 15<sup>th</sup> specimen of that species. In other words, we calculated the time it took to obtain 15 correctly identified specimens belonging to the same species.

In summary, the methods described above allowed us to measure two time lags (Fig 1):

- **Initial discovery, time lag 1.** The time lag between the first specimen collected of a species and the publication of the species name, calculated for all *Aframomum* species.

We also calculated the mean time lag ( $\pm$ SD) for *Aframomum* and for geographically widespread, medium or restricted species in this genus.

- **Extended discovery, time lag 2.** The time lag between the date the first specimen was collected and the year in which 15 specimens were collected and correctly determined, calculated for each species. We also calculated the mean time lag ( $\pm$ SD) for all available *Aframomum* species and for geographically widespread, medium or restricted *Aframomum* species.

### Time lags in other groups of plants

In addition, we wanted to measure the time lags for groups of plants with different life-history traits, distribution patterns and different numbers of species from those in *Aframomum* to allow comparison. We thus obtained data for conifers, which are mainly temperate (Dataset 3); for the tree genus *Leucaena* Benth. (Dataset 4); and for the family Acanthaceae Juss., which mainly has a tropical distribution but contains many more species than *Aframomum* (Dataset 5). The conifer dataset includes all species recorded by the authors of *An atlas of the world's conifers* (Farjon & Filer, 2013), who cited specimens that were useful for generating an accurate map for each species and therefore representative of their geographical distribution. Dataset 4 includes all records associated with the publication of a monograph of *Leucaena* (Hughes, 1998). Finally, Dataset 5 includes all GBIF occurrence records for the tropical family Acanthaceae with a collection date and associated with preserved specimens (2,492 names in total).

Finally, in a later stage we wanted to determine whether our results for *Aframomum*, conifers, *Leucaena* and Acanthaceae can be generalized to more plant taxa. We specifically used GBIF data to measure the time lag associated with obtaining the first 15 specimens of all species in the 20 largest families of angiosperms according to the World Checklist of Selected Plant Families—all of them with a large number of species with a tropical distribution (Royal

1  
2  
3 141 Botanic Gardens, Kew, 2019). This dataset (Dataset 6) includes all GBIF records with a  
4  
5 142 collection date and associated with preserved specimens: 164,598 species names associated  
6  
7 143 with 22,272,833 records.  
8  
9

10 144 For datasets 3 to 6, we extracted the earliest collection date, the date of the 15<sup>th</sup> specimen  
11  
12 145 and the total number of specimens for all taxa. For *Leucaena*, Acanthaceae and conifers  
13  
14 146 (Datasets 3, 4 and 5 respectively), we calculated the average time lag between the first and the  
15  
16 147 15<sup>th</sup> specimen for 586, 28 and 2492 taxa respectively. For the largest families of flowering  
17  
18 148 plants (Dataset 6), we calculated the average time lag between the first and the 15<sup>th</sup> specimen  
19  
20 149 for the 82,974 taxa in the twenty families that met this requirement. In all cases, we only  
21  
22 150 included records associated with preserved specimens—for example observation records were  
23  
24 151 excluded.  
25  
26  
27

28 152 **Additional time lag measurements**  
29

30 153 We extracted some additional information from Acanthaceae and the largest flowering  
31  
32 154 plant families for subsequent comparisons. First, we also measured how many names had their  
33  
34 155 first specimen collected after 1950 (roughly the last 70 years) and how many of these have  
35  
36 156 subsequently accumulated 15 or more specimens (Table 1). This allowed us to estimate the  
37  
38 157 impact of the previously reported increase in herbarium collections in the last decades  
39  
40 158 (Goodwin et al., 2015). Secondly, we calculated how long it took to collect as little as the first  
41  
42 159 three specimens for all taxa in these two datasets. This allowed us to compare our results with  
43  
44 160 IUCN conservation assessment guidelines (IUCN, 2016) and with a recent study on species  
45  
46 161 rarity (Enquist et al., 2019).  
47  
48  
49  
50

51 162 **Results**  
52

53 163 **Specimens cited at the time of publication**  
54

55 164 A total of 3,305 new seed plant species were published between 1970 and 2010 in *Kew*  
56  
57 165 *Bulletin*. The number of specimens cited per species in the protologue range from 1 to 155 with  
58  
59  
60



a mean of 4.9 ( $\pm 7.06$  SD). This distribution is a power law or hollow curve (Dial & Marzluff, 1989; Willis, 1922) (Fig 2A) in that 89% of protologues cite fewer than 10 specimens. In addition, only 6.2% of protologues cite 15 or more specimens (Fig 2B).

### Time lags in *Aframomum*

The average time between the first specimen collection and the year of publication (time lag 1) for all *Aframomum* species is 40.8 years ( $\pm 40.1$ , N = 59; Fig 1 and Table S4). Geographically widespread *Aframomum* species have a shorter lag ( $36.7 \pm 55.6$ , N = 9) than those with medium ( $46.2 \pm 40.5$ , N = 14) or restricted ( $39.9 \pm 35.2$ , N = 36) distributions.

The time lag between the collection of the first specimen and the collection of the first 15 specimens is 65.0 years ( $\pm 33.8$ ; Figs 1 and 3 and Table S5). However, the average time between the collection of the first specimen and the *correct* determination of 15 specimens is 100.8 years ( $\pm 38.8$ , N = 29; Figs 1 and 3). In other words, the time taken to obtain 15 correctly identified specimens added an additional 35.8 years to the collection of 15 specimens (Fig 3).

Geographically restricted *Aframomum* species have a shorter lag to the correct determination of 15 specimens ( $75.8 \pm 44.1$ , N = 4) than species with medium ( $148.3 \pm 61.4$ , N = 14) or widespread ( $145.0 \pm 61.4$ , N = 9) distributions, although the sample size is too small to be statistically significant.

### Time lags in other groups of plants

The average time between the collection of the first specimen and the collection of the 15<sup>th</sup> specimen (time lag 2) is 68.6 years ( $\pm 35.8$ , N = 28) in *Leucaena*, 87.4 years ( $\pm 49.0$ , N = 586) in conifers and 71.1 years ( $\pm 43.9$ , N = 2,492) in Acanthaceae (Fig 4A and Table 1). Average time in the twenty largest families of flowering plants ranges from  $58.6 \pm 41.9$  years in Arecaceae to  $74.3 \pm 52.9$  in Lamiaceae (Fig 4B and Table 1).

### Time lags to collect three specimens

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

The average time lag for the collection of only three specimens in the different groups is notably lower but still significant: 35.72 ( $\pm 35.03$ ,  $N = 58$ ) in *Aframomum* and 33.76 ( $\pm 34.20$ ,  $N = 4,203$ ) in Acanthaceae (Fig 5). Similar values are seen in the largest families of Angiosperms, from 27.84  $\pm$  28.42 years in Arecaceae to 34.60  $\pm$  40.49 in Amaryllidaceae (Fig 6 and Data File 6).

**Discussion**

Even a basic understanding of species takes a long time to accumulate. We measured three key stages (Fig 1) and quantified the two time lags associated with accumulating knowledge of species of the recently-monographed genus *Aframomum* and other diverse groups of plants. First, the description of new species in *Aframomum* took on average four decades from the collection of their first specimen (time lag 1 in Fig 1). Furthermore, additional specimens up to the minimum threshold of 15 were collected over 25 years after the formal publication of the species name, but it took over six decades after publication to obtain 15 correctly identified specimens (time lag 2 in Fig 1). In summary, it took an average of one hundred years to gather a basic understanding of the species in *Aframomum*.

In line with our results for *Aframomum*, barely 5% of the 3,305 new species published over a 40-year period in *Kew Bulletin* cited 15 or more specimens (Fig 2B). On the contrary, new descriptions cited on average 4.9 specimens per species, far less than the 15 we consider reasonable for a minimum knowledge. It is therefore likely that the period of Extended discovery of these species, time lag 2, also extended for several decades after their formal publication.

Nevertheless, the extensive time lags we found for *Aframomum* led us to consider whether these were atypically long given that *Aframomum* is difficult to collect and identify. Subsequently, we examined a number of other data sets to determine whether they showed a different discovery timeframe. Our datasets include data with varying degrees of accuracy and

completeness. At one extreme, in the *Aframomum* dataset, the entire determination history is known for every specimen studied. At the other extreme, the upprocessed data from GBIF likely includes synonyms and misidentified specimens. Our reasoning was that because the time lags we found in *Aframomum* were so long, a similar signal may be likely to be found even for taxonomically unprocessed datasets such as these. The additional datasets are also less comprehensive than the *Aframomum* dataset, since we had no exact dates for the correct determination of each specimen. For this reason, in the analysis of these datasets we were only able to compare the time lags for the collection of 3 and 15 specimens. In addition, the datasets using GBIF data (datasets 5 and 6), albeit informative, are likely to be incomplete as not all herbarium specimens of these taxa have been made available through the GBIF portal. Finally, another caveat is that the dataset with the twenty largest families of plants includes many species from temperate regions. The inclusion of these species in our analyses—in principle better known and more frequently collected—likely results in mean estimates lower than those obtained for tropical taxa. In summary, results obtained from analysing these additional datasets should be carefully interpreted, taking into account their limitations and potential biases. Nevertheless, these datasets provide additional independent evidence that the patterns in *Aframomum* apply to other groups of plants.

The average number of years to collect 15 specimens is 87 for conifers, 68 for *Leucaena* and 71 for Acanthaceae, broadly similar to the 65 for *Aframomum* (Fig 4A and Table 1). Nevertheless, the fact that the time lags for the different taxa are broadly in agreement provides confidence that the reported results for *Aframomum* are applicable to other groups of plants. Furthermore, the same patterns are observed when the largest flowering plant families are analysed (Fig 4B). Species in these families required on average 68 years to acquire 15 specimens, and no less than three decades to acquire only 3 (Fig 6 and Data File 6). Considering

1  
2  
3 239 that many species in these families have a temperate distribution, which makes them likely to  
4  
5 240 be better collected, it is possible that the time lags for tropical species in this families are longer.  
6  
7  
8 241 The validity of our results is corroborated by results in other publications. The time lag  
9  
10 242 from collection until first publication of the species name (time lag 1), for example, is 40 years  
11  
12 243 for *Aframomum* as determined in this study, compared to a mean value of 38.8 years for species  
13  
14 244 in various flowering plant monographs and 32.4 for all species published in the *Kew Bulletin*  
15  
16  
17 245 over a 40-year period (Bebber et al., 2010). Consequently, we see no compelling reason why  
18  
19 246 time lag 2 would not have similar temporal dynamics and show similar trends in other genera  
20  
21 247 and families. In addition, the fact that very few plant protologues cite more than five specimens  
22  
23  
24 248 means that time lag 2 for all species must almost always be longer than lag 1. We have also  
25  
26 249 shown previously that extensive mis-identifications are found in other tropical plant taxa as  
27  
28 250 well as *Aframomum* when measured in a number of different ways (Goodwin et al., 2015).  
29  
30 251 Thus, although we assume the exact duration of time lag 2 —100.8 years in *Aframomum*—  
31  
32 252 will vary in other taxa, and for some particularly well-known groups will be shorter, we  
33  
34  
35 253 consider that our results clearly show that the time lag from the first specimen of a species  
36  
37 254 being collected to obtaining 15 correctly identified specimens of that species is a surprisingly  
38  
39  
40 255 long one.

41  
42 256 Finally, we also measured the extent that time lags 1 and 2 differ between species which  
43  
44 257 are restricted and widespread in distribution. The differences were relatively modest  
45  
46  
47 258 (Supplementary information 1) but we note that widespread *Aframomum* species have  
48  
49 259 relatively shorter time lags as they were collected more quickly than restricted species. Similar  
50  
51 260 patterns have been observed in other groups of organisms (Blackburn & Gaston, 1995; Gaston  
52  
53 261 et al., 1995; Gaston & Blackburn, 1994).

54  
55  
56 262 One caveat with the *Aframomum* data used in this analysis is that it is a subset  
57  
58 263 (approximately 50%) of all *Aframomum* herbarium specimens; only 1,779 specimens had a  
59  
60

complete set of dates for every determination. Of this subset, only species with a minimum of 15 specimens in the analysed subset could be investigated for lag 2. The main consequence of this is that the dates for the collection and correct determination of 15 specimens represents the latest possible dates that these key stages were completed. However, the subset of species with a minimum of 15 specimens analysed have a time lag 1 comparable to that of all *Aframomum* species (Table S4), suggesting other time lags will also be similar. Additionally, the lags between the collection of the first specimen and the collection of 15 specimens for the complete datasets 3–6 are remarkably similar to that of *Aframomum*.

### How long does it take to know a species?

“Knowing” a species is an ongoing process, from the first collection to an indepth understanding of its frequency and geographical distribution pattern, alongside other aspects such as its ecology or breeding behaviour. In this sense, the choice to use 15 accurately determined specimens to calculate our lag periods is arbitrary yet serves as a useful heuristic proxy in demonstrating the long time lags associated with the discovery of new species. We consider that our results are broadly compatible with three distinct phases of knowing a species and we suggest that this is a useful framework with which to understand species discovery (Fig 1). This framework begins with the **initial discovery** phase (time lag 1), which runs from the collection of the first specimen until the publication of the species as new to science. The **extension discovery** phase (time lag 2) follows the initial discovery phase and includes the collection of enough correctly determined specimens to provide the ground for any subsequent study. After the initial and extension discovery phases, comes the **consolidation** phase where a species is known in greater detail. This is not to say that a “eureka” moment does not exist; there is definitely a moment when a botanist realises that they are looking at a new, undescribed species, which leads to the publication of a new species name (time lag 1). However, we argue that this *recognition* of a new taxa is not the same as *knowing* the new taxa. In fact, the majority

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

of tropical species remain somewhere in the initial and extension discovery phases. The current immature consolidation phase for tropical species contrasts sharply with that of nearly all temperate floras and it is difficult to imagine that anything other than increased capacity in taxonomic botany can realistically address this shortfall in our knowledge of tropical plants.

**Conclusion**

Species of flowering plants have been named at a relatively uniform rate of approximately 1,500 - 2,000 species per year since at least 1970, with approximately 100,000 species described over the last five decades (Bebber et al., 2014). Given the results reported here, we must assume that many of these species are still in a relatively early stage of being known at a sufficient level to infer a reasonable distribution map. In line with this, the analysis of Dataset 6 in this study —GBIF data for all species in the twenty largest flowering plant families— shows that only 50% of species (82,974 out of 164,958) have accumulated at least 15 records, and that only 18.5% of species first collected after 1950 have 15 or more specimens. Even proceeding with caution due to the fragmentary nature of these data, our results suggest that the doubling of herbarium specimens in the world’s herbaria in the last sixty years (Goodwin et al., 2015) has not necessarily led to an uniform increase in specimens for all species.

The significance of the long period required to obtain an adequate account of a newly discovered species is particularly apparent when considered in its stark contrast to the rates of predicted and already-documented change negatively impacting global biodiversity. Our analyses provide a timeframe with implications for other activities that utilize herbarium data such as monitoring global biodiversity, making informed conservation assessments, constructing accurate distribution maps and measuring extinction rates.

**Funding**

This research was funded by an NERC studentship to ZAG and a Leverhulme Trust and SynTax awards from BBSRC, NERC, Systematics Association and the Linnean Society to RWS.

## Author contribution

RWS & DJH supervised the project. ZAG and PMR collected the data and performed the analyses. ZAG, PMR, DJH, TW, JRIW, DF and RWS interpreted the results. ZAG, PMR, TW and RWS wrote the paper.

## Data availability

All processed datasets and scripts used in this study are available as supplementary material. Unprocessed *Aframomum*, *Leucaena* and conifer data are also available as supplementary material, whereas unprocessed GBIF datasets are available at the following DOI: Acanthaceae (<https://doi.org/10.15468/dl.j1755h>), Amaryllidaceae (<https://doi.org/10.15468/dl.1r6zcf>), Apocynaceae (<https://doi.org/10.15468/dl.un40hk>), Araceae (<https://doi.org/10.15468/dl.lfybyz>), Areaceae (<https://doi.org/10.15468/dl.qvbenq>), Asparagaceae (<https://doi.org/10.15468/dl.xw1xqq>), Asteraceae (<https://doi.org/10.15468/dl.ptcozj>), Begoniaceae (<https://doi.org/10.15468/dl.sl4czw>), Bromeliaceae (<https://doi.org/10.15468/dl.igfnvd>), Campanulaceae (<https://doi.org/10.15468/dl.nwfzmi>), Convolvulaceae (<https://doi.org/10.15468/dl.d5pvuc>), Cyperaceae (<https://doi.org/10.15468/dl.bos5tn>), Euphorbiaceae (<https://doi.org/10.15468/dl.connvt>), Fabaceae (<https://doi.org/10.15468/dl.9wnxau>), Iridaceae (<https://doi.org/10.15468/dl.mixagg>) Lamiaceae (<https://doi.org/10.15468/dl.jc1fct>), Myrtaceae (<https://doi.org/10.15468/dl.ftwvex>), Orchidaceae (<https://doi.org/10.15468/dl.ytdnab>), Phyllanthaceae (<https://doi.org/10.15468/dl.avfmhi>), Poaceae (<https://doi.org/10.15468/dl.grswvm>), Rubiaceae (<https://doi.org/10.15468/dl.oh9qab>). The dataset by Harris and Wotley (2018) for their

338 monographic study of *Aframomum* is available at Dryad repository

339 (<https://doi.org/10.5061/dryad.03622>).

## 340 References

- 341 Bebber, D. P., Carine, M. A., Wood, J. R. I., Wortley, A. H., Harris, D. J., Prance, G. T.,  
342 Davidse, G., Paige, J., Pennington, T. D., Robson, N. K. B. & Scotland, R. W. (2010).  
343 Herbaria are a major frontier for species discovery. *PNAS*, *107*, 22169–22171.
- 344 Bebber, D. P., Wood, J. R. I., Barker, C. & Scotland, R. W. (2014). Author inflation masks  
345 global capacity for species discovery in flowering plants. *New Phytologist*, *201*, 700–  
346 706. <https://doi.org/10.1111/nph.12522>
- 347 Blackburn, T. M. & Gaston, K. J. (1995). What determines the probability of discovering  
348 species?: A study of South American oscine passerine birds. *Journal of*  
349 *Biogeography*, *22*, 7–14.
- 350 Castañeda-Álvarez, N. P., Khoury, C. K., Achicanoy, H. A., Bernau, V., Dempewolf, H.,  
351 Eastwood, R. J., Guarino, L., Harker, R. H., Jarvis, A., Maxted, N., Müller, J. V.,  
352 Ramirez-Villegas, J., Sosa, C. C., Struik, P. C., Vincent, H. & Toll, J. (2016). Global  
353 conservation priorities for crop wild relatives. *Nature Plants*, *2*, 16022.  
354 <https://doi.org/10.1038/nplants.2016.22>
- 355 CBD (2010). *X/2.Strategic Plan for Biodiversity 2011-2020*. Convention on Biological  
356 Diversity
- 357 Costello, M. J., May, R. M. & Stork, N. E. (2013). Can we name Earth's species before they  
358 go extinct? *Science*, *339*, 413–416. <https://doi.org/10.1126/science.1230318>
- 359 Dial, K. P. & Marzluff, J. M. (1989). Nonrandom diversification within taxonomic  
360 assemblages. *Systematic Biology*, *38*, 26–37. <https://doi.org/10.1093/sysbio/38.1.26>
- 361 Enquist, B. J., Feng, X., Boyle, B., Maitner, B., Newman, E. A., Jørgensen, P. M.,  
362 Roehrdanz, P. R., Thiers, B. M., Burger, J. R., Corlett, R. T., Couvreur, T. L. P.,  
363 Dauby, G., Donoghue, J. C., Foden, W., Lovett, J. C., Marquet, P. A., Merow, C.,  
364 Midgley, G., Morueta-Holme, N., Neves, D. M., Oliveira-Filho, A. T., Kraft, N. J. B.,  
365 Park, D. S., Peet, R. K., Pillet, M., Serra-Diaz, J. M., Sandel, B., Schildhauer, M.,  
366 Šimová, I., Violle, C., Wieringa, J. J., Wiser, S. K., Hannah, L., Svenning, J.-C. &  
367 McGill, B. J. (2019). The commonness of rarity: global and future distribution of  
368 rarity across land plants. *Science Advances*, *5*, eaaz0414.  
369 <https://doi.org/10.1126/sciadv.aaz0414>
- 370 Farjon, A. & Filer, D. (2013). *An Atlas of the World's Conifers*. BRILL
- 371 Fontaine, B., Perrard, A. & Bouchet, P. (2012). 21 years of shelf life between discovery and  
372 description of new species. *Current Biology*, *22*, R943–R944.  
373 <https://doi.org/10.1016/j.cub.2012.10.029>
- 374 Gaston, K. J. & Blackburn, T. M. (1994). Are newly described bird species small-bodied?  
375 *Biodiversity Letters*, *2*, 16–20.



- 376 Gaston, K. J., Blackburn, T. M. & Loder, N. (1995). Which species are described first?: The  
377 case of North American butterflies. *Biodiversity and Conservation*, 4, 119–127.
- 378 Goodwin, Z. A., Harris, D. J., Filer, D., Wood, J. R. I. & Scotland, R. W. (2015). Widespread  
379 mistaken identity in tropical plant collections. *Current Biology*, 25, R1066–R1067.  
380 <https://doi.org/10.1016/j.cub.2015.10.002>
- 381 Harris, D. J. & Wortley, A. H. (2018). *Monograph of Aframomum (Zingiberaceae)*. The  
382 American Society of Plant Taxonomists, Laramie WY 82071, USA
- 383 Hawthorne, W. D. & Marshall, C. A. M. (2016). A manual for Rapid Botanic Survey (RBS)  
384 and measurement of vegetation bioquality.
- 385 Hughes, C. (1998). *Monograph of Leucaena (Leguminosae-Mimosoideae)*. The American  
386 Society of Plant Taxonomists, Ann Arbor. 244 pp.
- 387 IUCN (2016). *Guidelines for using the IUCN Red List categories and criteria. Version 12*.  
388 Standards and Petitions Subcommittee, IUCN
- 389 Kraft, C. S., McAdam, A. J. & Carroll, K. C. (2017). A rose by any other name: practical  
390 updates on microbial nomenclature for clinical microbiology. C.-A. D. Burnham (ed.).  
391 *Journal of Clinical Microbiology*, 55, 3–4. <https://doi.org/10.1128/JCM.02169-16>
- 392 Mace, G. M. (2004). The role of taxonomy in species conservation. H. C. J. Godfray & S.  
393 Knapp (eds.). *Philosophical Transactions of the Royal Society of London. Series B:*  
394 *Biological Sciences*, 359, 711–719. <https://doi.org/10.1098/rstb.2003.1454>
- 395 Nic Lughadha, E. M., Grazielle Staggemeier, V., Vasconcelos, T. N. C., Walker, B. E.,  
396 Canteiro, C. & Lucas, E. J. (2019). Harnessing the potential of integrated systematics  
397 for conservation of taxonomically complex, megadiverse plant groups. *Conservation*  
398 *Biology*, 33, 511–522. <https://doi.org/10.1111/cobi.13289>
- 399 Raven, P. H. (2007). Foreword. In: C. Jarvis (ed.) *Order out of chaos: Linnaean plant names*  
400 *and their types*. Linnean Society of London & Natural History Museum, London,  
401 London, p. 1024.
- 402 Royal Botanic Gardens, Kew (2019). World Checklist of Selected Plant Families. *World*  
403 *Checklist of Selected Plant Families*. Available from: <http://wcp.science.kew.org/>  
404 (August 1, 2019)
- 405 Steffen, W., Richardson, K., Rockstrom, J., Cornell, S. E., Fetzer, I., Bennett, E. M., Biggs,  
406 R., Carpenter, S. R., de Vries, W., de Wit, C. A., Folke, C., Gerten, D., Heinke, J.,  
407 Mace, G. M., Persson, L. M., Ramanathan, V., Reyers, B. & Sorlin, S. (2015).  
408 Planetary boundaries: guiding human development on a changing planet. *Science*,  
409 347, 1259855–1259855. <https://doi.org/10.1126/science.1259855>
- 410 Willis, J. C. (1922). *Age and area: a study in geographical distribution and origin of species*  
411 *(classic reprint)*. Forgotten Books, Place of publication not identified
- 412

**Table 1 Extended discovery time lags across plant groups.** Average (mean) time in years between the collection of the first specimen and the collection of 15 specimens in *Aframomum*, conifers, *Leucaena*, Acanthaceae and the twenty largest families of flowering plants.

Group	Number of species	Average time in years (± SD) until 15 specimens are available
<i>Aframomum</i>	29	65.0 ± 33.8
conifers	586	87.4 ± 49.0
<i>Leucaena</i>	28	68.6 ± 35.8
Acanthaceae	2,492	71.1 ± 43.9
Asteraceae	c. 32,581*	67.4 ± 48.4
Orchidaceae	28,484*	70.3 ± 40.7
Fabaceae	20,856*	66.8 ± 45.9
Rubiaceae	13,765*	65.1 ± 43.7
Poaceae	11,467*	66.0 ± 43.1
Lamiaceae	7,587*	74.3 ± 52.9
Euphorbiaceae	6,462*	67.0 ± 45.5
Apocynaceae	6,314*	65.6 ± 42.2
Myrtaceae	6,019*	66.9 ± 49.0
Cyperaceae	5,539*	68.5 ± 42.3
Araceae	3,720*	62.9 ± 41.3
Bromeliaceae	3,465*	70.5 ± 41.9
Asparagaceae	3,102*	65.4 ± 44.8
Arecaceae	2,551*	58.6 ± 41.9
Iridaceae	2,481*	66.8 ± 40.2
Campanulaceae	2,419*	67.5 ± 46.7
Amaryllidaceae	2,140*	76.3 ± 54.3
Phyllanthaceae	2,050*	61.6 ± 41.6
Convolvulaceae	1,921*	73.3 ± 46.5
Begoniaceae	1,794†	70.9 ± 41.8

\* WCSP Kew. † Begonia database.

**Fig 1.** How long does it take to ‘know’ a species? The first herbarium specimen of the species is collected (key stage 1). It then takes, on average, several decades (time lag 1) for the new species to be formally described and published (key stage 2). Over time (time lag 2) fifteen specimens of the species are collected and determined correctly (key stage 3) allowing the production of a conservation assessment. Eventually these species are known in such a detail that monitoring would be possible. As an example, numbers on the bottom indicate the length of time lags 1 and 2 in *Aframomum*.

**Fig 2.** (A) Numbers of specimens cited per new species of seed plant published in *Kew Bulletin*, between 1970 and 2010 inclusive (N = 3,305). (B) Only 6.2% species had 15 or more specimens recorded at the time of publication, whereas 93.8% of them had less than 15 specimen records at the time of species description.

**Fig 3.** Time lag between the 1<sup>st</sup> specimen collected, the 15<sup>th</sup> specimen collected and the moment when at least 15 specimens are correctly determined for *Aframomum*. Each line represents a different species in the genus. Average time lag to accumulate 15 collections of *Aframomum* species is 60.5 years, whereas it takes, on average, another four decades to accumulate 15 correctly identified ones.

**Fig 4.** Time lag between (A) the 1<sup>st</sup> and the 15<sup>th</sup> specimen collected for all species in *Aframomum*, *Leucaena*, conifers and Acanthaceae, and (B) between the 1<sup>st</sup> and the 15<sup>th</sup> specimen collected for all species from the 20 largest families of flowering plants. Values in the graph represent taxa for each taxonomic group, with dots indicating outliers. The × symbols indicate mean values for each family.

**Fig 5** Time lags between the collection of the 1<sup>st</sup> and the 3<sup>rd</sup> and the 1<sup>st</sup> and the 15<sup>th</sup> specimens in *Aframomum* and Acanthaceae. Values in the graph represent taxa for each taxonomic group, with dots indicating outliers. The × symbols indicate mean values for each family.

**Fig 6** Time lag between (a) the 1<sup>st</sup> and the 3<sup>rd</sup> specimen collected for all species in the 20 largest families of flowering plants. Values in the graph represent taxa for each taxonomic group, with dots indicating outliers. The × symbols indicate mean values for each family.

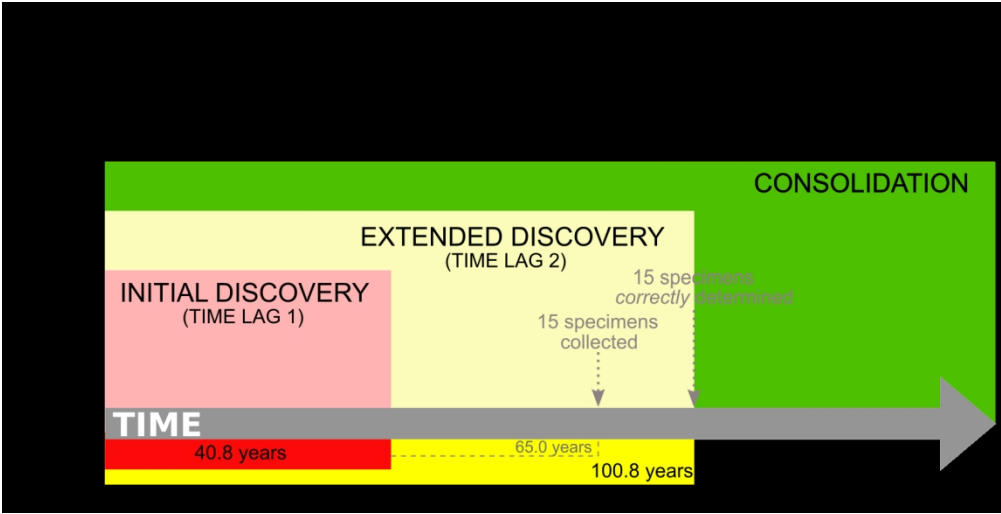


Fig 1. How long does it take to 'know' a species? The first herbarium specimen of the species is collected (key stage 1). It then takes, on average, several decades (time lag 1) for the new species to be formally described and published (key stage 2). Over time (time lag 2) fifteen specimens of the species are collected and determined correctly (key stage 3) allowing the production of a conservation assessment. Eventually these species are known in such a detail that monitoring would be possible. As an example, numbers on the bottom indicate the length of time lags 1 and 2 in *Aframomum*.

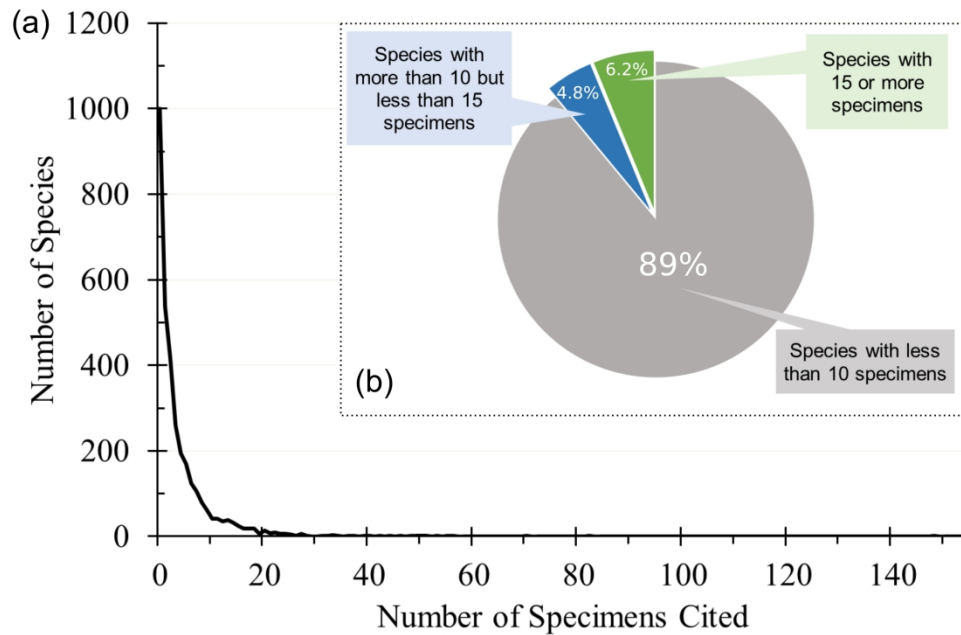


Fig 2. (A) Numbers of specimens cited per new species of seed plant published in Kew Bulletin, between 1970 and 2010 inclusive ( $N = 3,305$ ). (B) Only 6.2% species had 15 or more specimens recorded at the time of publication, whereas 93.8% of them had less than 15 specimen records at the time of species description.

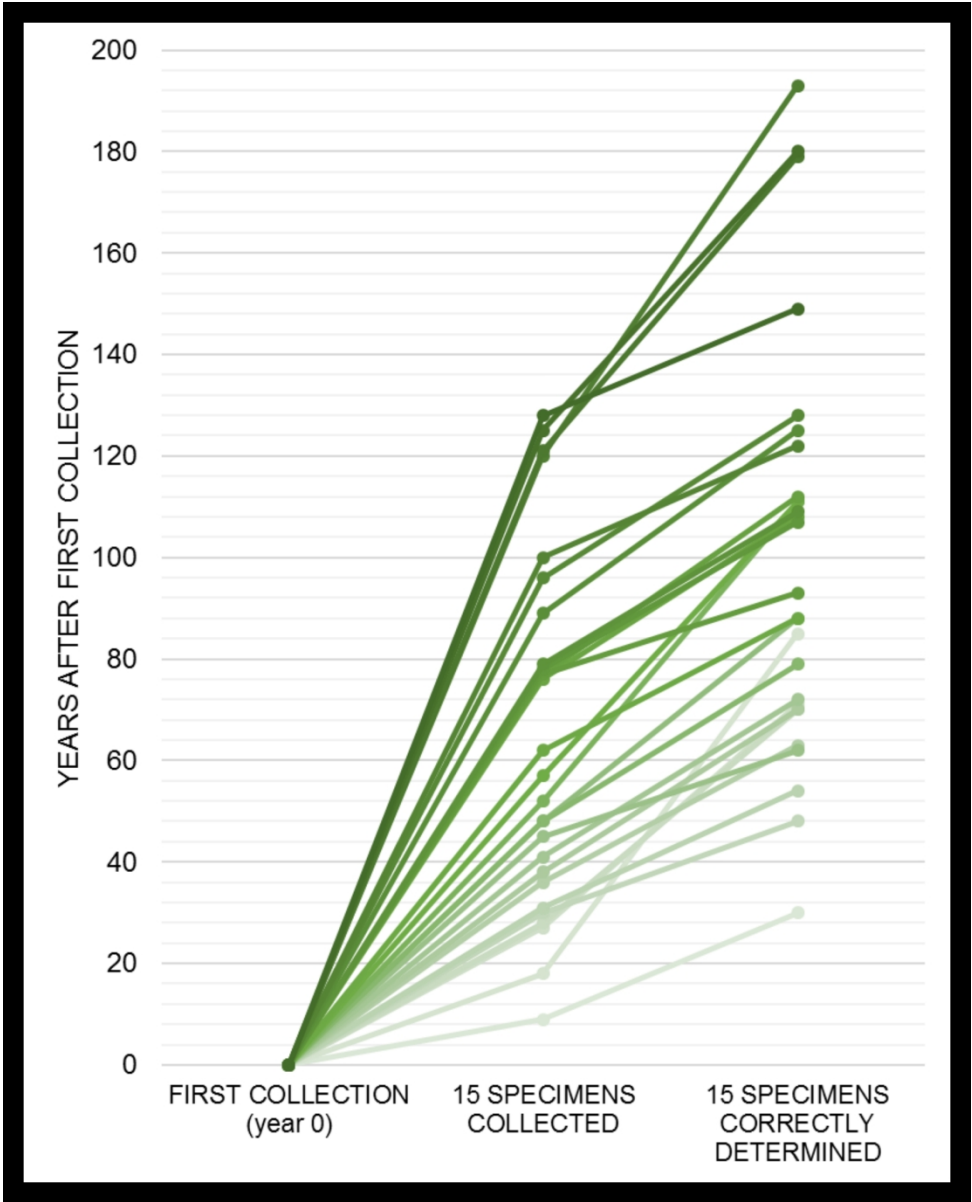


Fig. 3. Time lag between the 1st specimen collected, the 15th specimen collected and the moment when at least 15 specimens are correctly determined for *Aframomum*. Each line represents a different species in the genus. Average time lag to accumulate 15 of *Aframomum* species is 60.5 years, whereas it takes, on average, another four decades to accumulate 15 correctly identified ones.

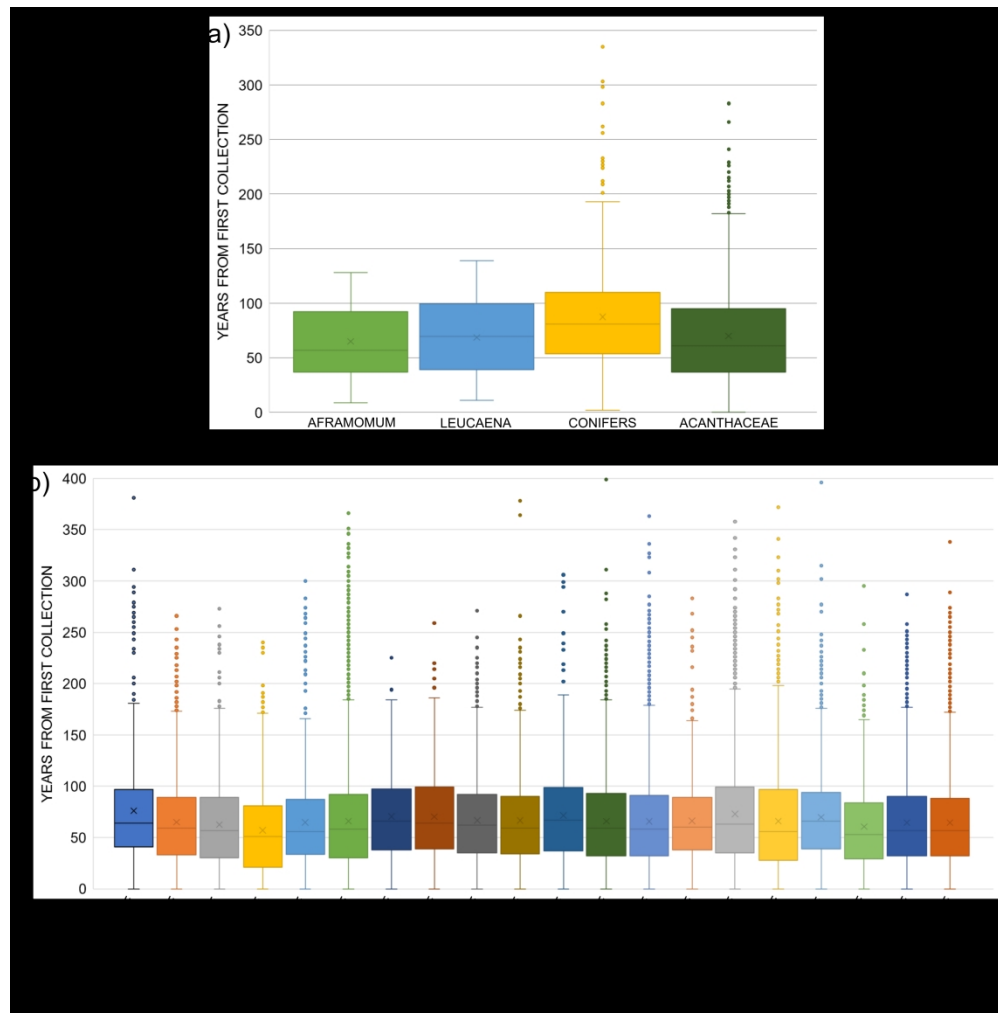


Fig 4. Time lag between (A) the 1st and the 15th specimen collected for all species in Aframomum, Leucaena, conifers and Acanthaceae, and (B) between the 1st and the 15th specimen collected for all species from the 20 largest families of flowering plants. Values in the graph represent taxa for each taxonomic group, with dots indicating outliers. The × symbols indicate mean values for each family.

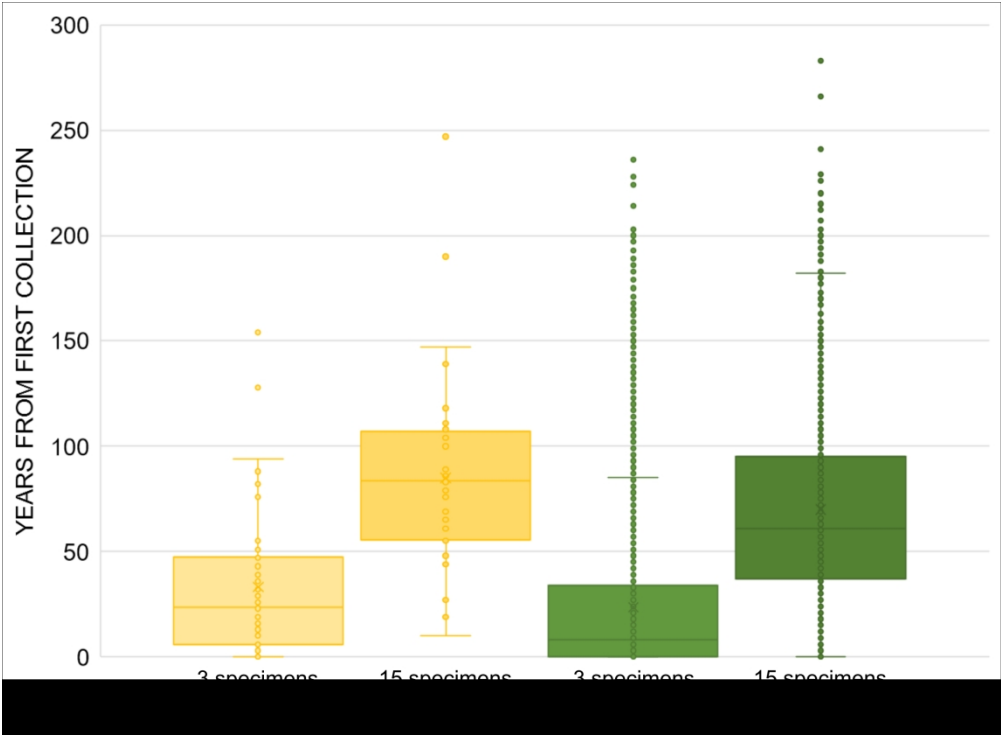


Fig 5 Time lags between the collection of the 1st and the 3rd and the 1st and the 15th specimens in Aframomum and Acanthaceae. Values in the graph represent taxa for each taxonomic group, with dots indicating outliers. The × symbols indicate mean values for each family.



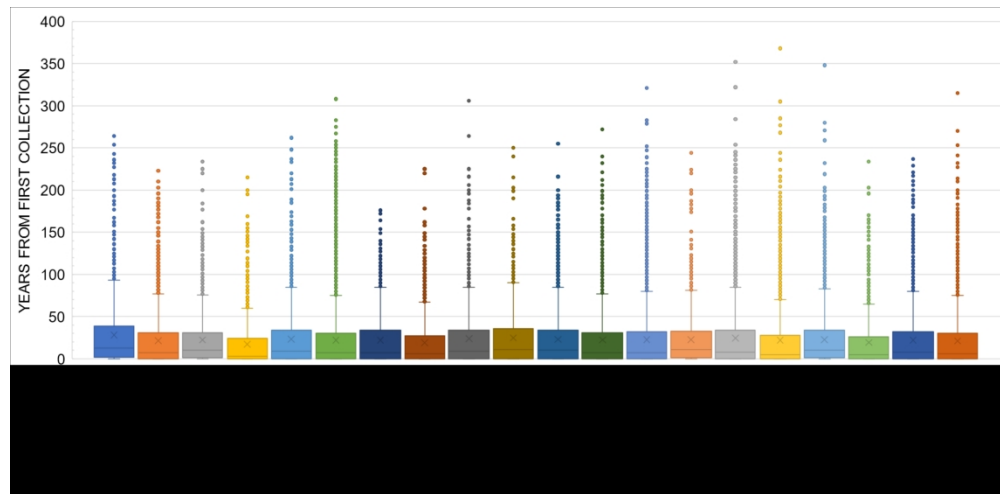


Fig 6 Time lag between (a) the 1st and the 3rd specimen collected for all species in the 20 largest families of flowering plants. Values in the graph represent taxa for each taxonomic group, with dots indicating outliers. The × symbols indicate mean values for each family.

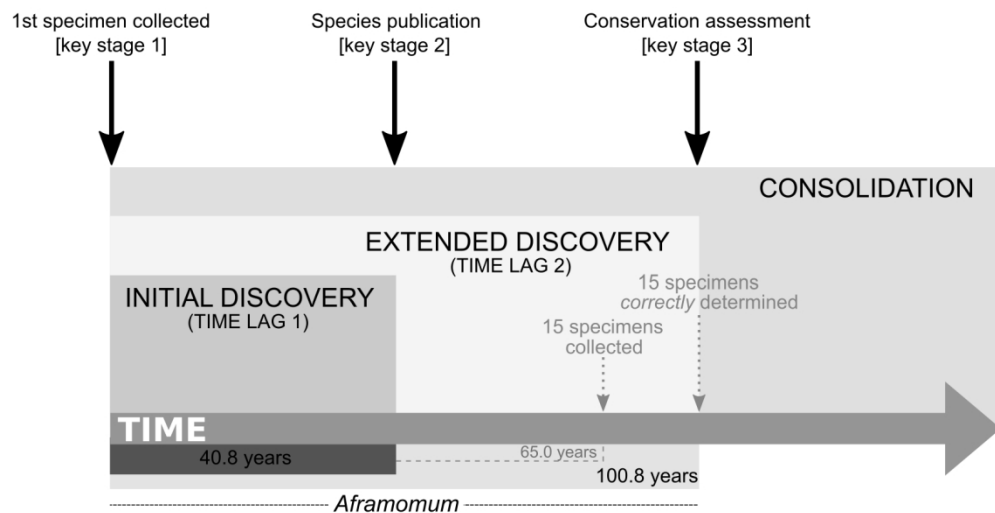


Fig 1. How long does it take to 'know' a species? The first herbarium specimen of the species is collected (key stage 1). It then takes, on average, several decades (time lag 1) for the new species to be formally described and published (key stage 2). Over time (time lag 2) fifteen specimens of the species are collected and determined correctly (key stage 3) allowing the production of a conservation assessment. Eventually these species are known in such a detail that monitoring would be possible. As an example, numbers on the bottom indicate the length of time lags 1 and 2 in *Aframomum*.

215x109mm (200 x 200 DPI)

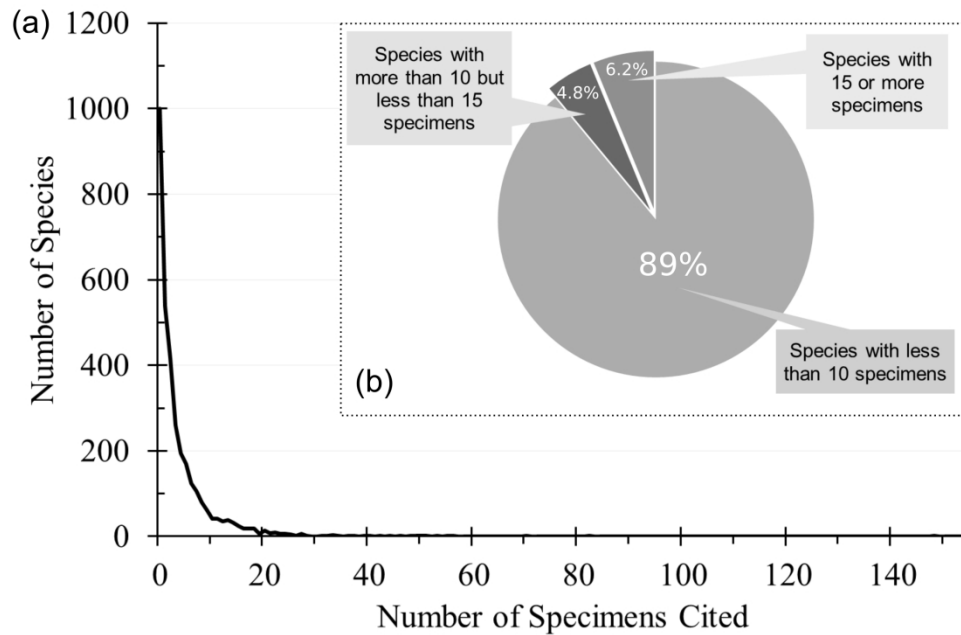
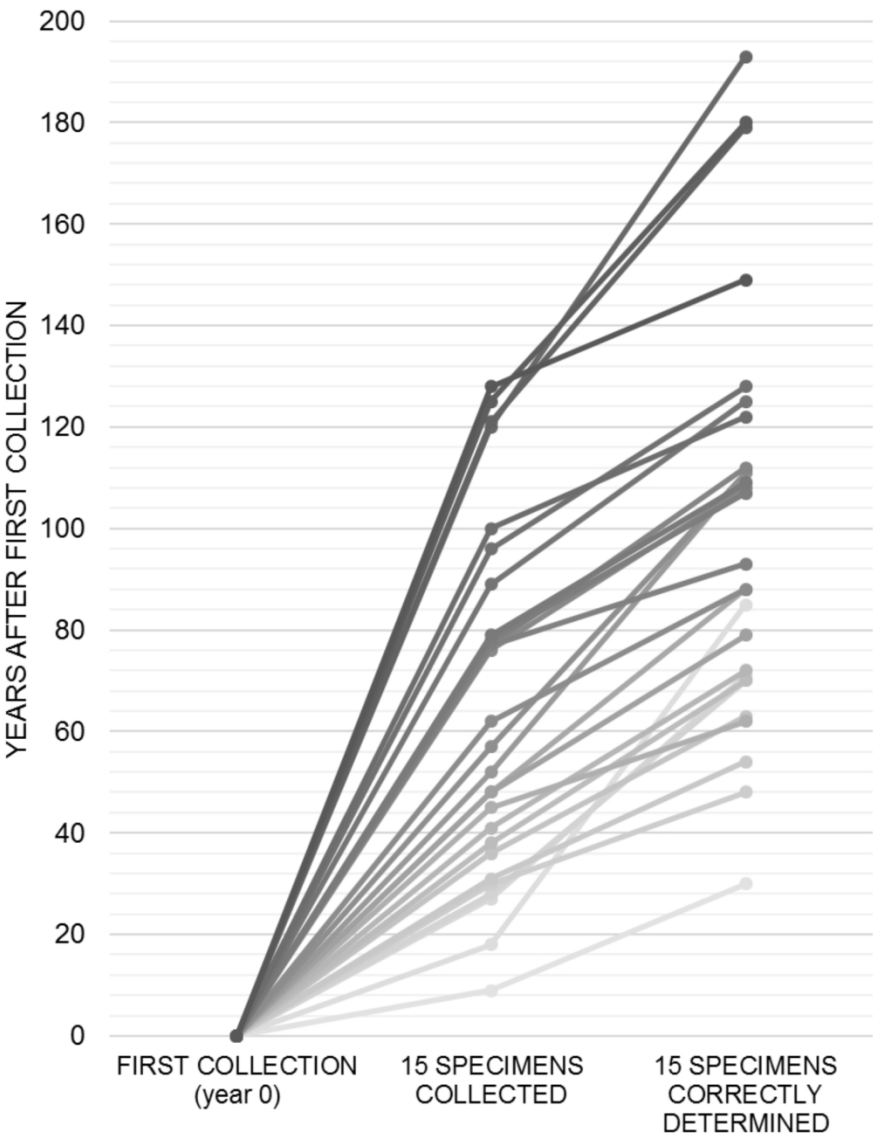


Fig 2. (A) Numbers of specimens cited per new species of seed plant published in Kew Bulletin, between 1970 and 2010 inclusive ( $N = 3,305$ ). (B) Only 6.2% species had 15 or more specimens recorded at the time of publication, whereas 93.8% of them had less than 15 specimen records at the time of species description.

335x218mm (200 x 200 DPI)



232x287mm (200 x 200 DPI)

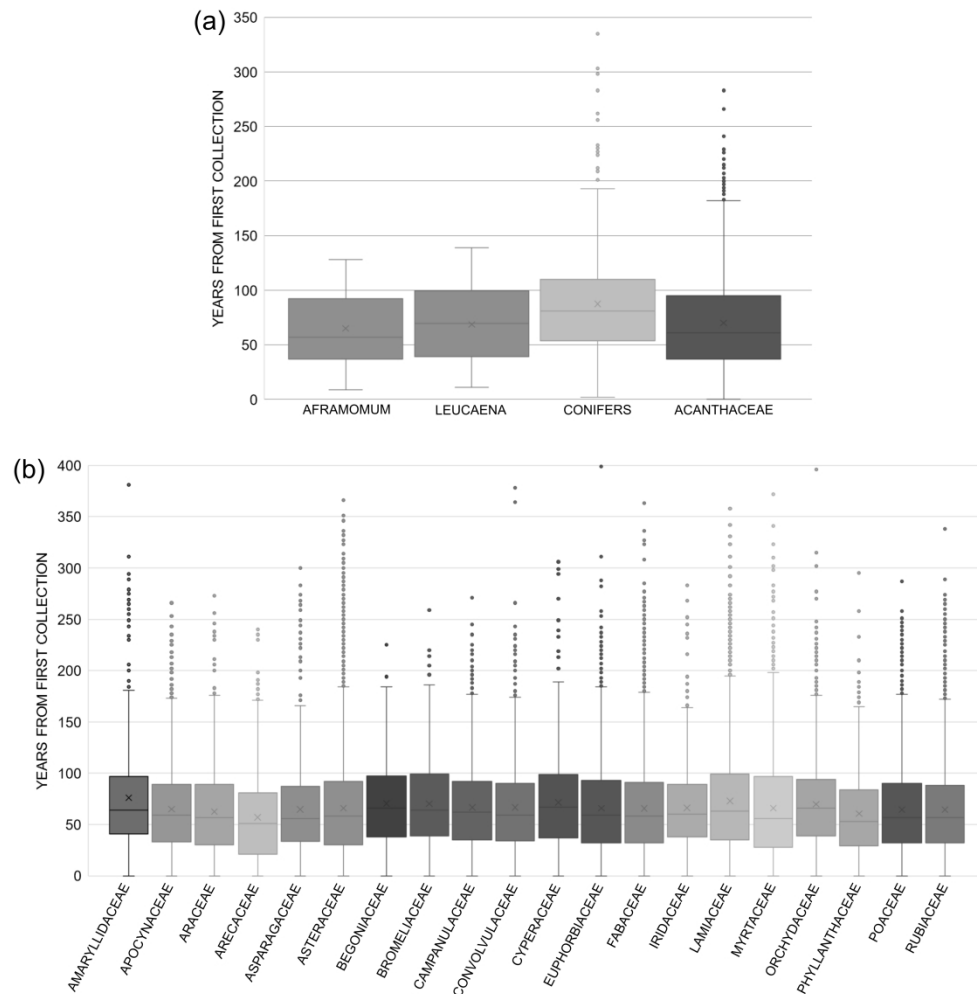


Fig 4. Time lag between (A) the 1st and the 15th specimen collected for all species in Aframomum, Leucaena, conifers and Acanthaceae, and (B) between the 1st and the 15th specimen collected for all species from the 20 largest families of flowering plants. Values in the graph represent taxa for each taxonomic group, with dots indicating outliers. The × symbols indicate mean values for each family.

534x541mm (200 x 200 DPI)

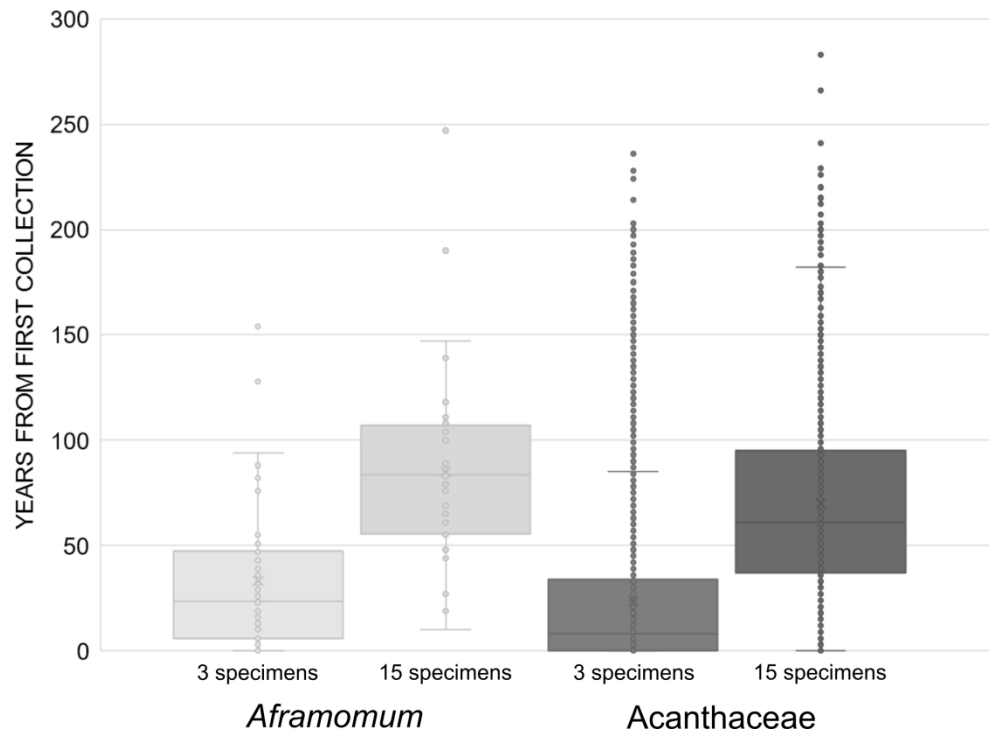


Fig 5 Time lags between the collection of the 1st and the 3rd and the 1st and the 15th specimens in *Aframomum* and *Acanthaceae*. Values in the graph represent taxa for each taxonomic group, with dots indicating outliers. The x symbols indicate mean values for each family.

326x239mm (200 x 200 DPI)

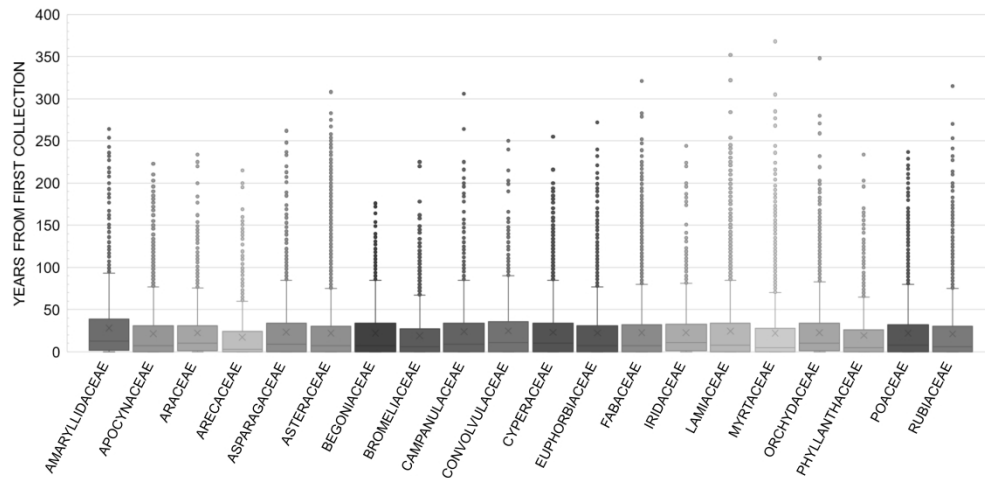


Fig 6 Time lag between (a) the 1st and the 3rd specimen collected for all species in the 20 largest families of flowering plants. Values in the graph represent taxa for each taxonomic group, with dots indicating outliers. The × symbols indicate mean values for each family.

518x254mm (200 x 200 DPI)