Herbaria are a major frontier for species discovery

Daniel P. Bebber[a], Mark A. Carine[b], John R. I. Wood[c], Alexandra H. Wortley[d], David J. Harris[d], Ghillean T. Prance[e], Gerrit Davidse[f], Jay Paige[f], Terry D. Pennington[e], Norman K. B. Robson[b], and Robert W. Scotland[c,1]

[a]Earthwatch Institute, Oxford OX2 7DE, United Kingdom; [b]Department of Botany, Natural History Museum, London SW7 5BD, United Kingdom; [c]Department of Plant Sciences, University of Oxford, Oxford OX1 3RB, United Kingdom; [d]Royal Botanic Garden Edinburgh, Edinburgh EH3 5LR, United Kingdom; [e]Royal Botanic Gardens, Kew, Richmond, Surrey TW93AB, United Kingdom; and [f]Missouri Botanical Garden, St. Louis, MO 63166-0299

Edited by David B. Wake, University of California, Berkeley, CA, and approved November 2, 2010 (received for review August 11, 2010)

Despite the importance of species discovery, the processes including collecting, recognizing, and describing new species are poorly understood. Data are presented for flowering plants, measuring quantitatively the lag between the date a specimen of a new species was collected for the first time and when it was subsequently described and published. The data from our sample of new species published between 1970 and 2010 show that only 16% were described within five years of being collected for the first time. The description of the remaining 84% involved much older specimens, with nearly one-quarter of new species descriptions involving specimens >50 y old. Extrapolation of these results suggest that, of the estimated 70,000 species still to be described, more than half already have been collected and are stored in herbaria. Effort, funding, and research focus should, therefore, be directed as much to examining extant herbarium material as collecting new material in the field.

herbarium specimen | monograph | taxonomy

Accurate species recognition underpins our knowledge of global biodiversity (1–3). In recent years, the lack of taxonomic activity has led to increased political (4) and scientific calls (3) to invest in the science of taxonomy, which is fundamental for what we know about species-level diversity. The assumptions behind these demands are that increased resources would necessarily lead to increased taxonomic productivity and accuracy. Given finite resources, it is essential that scientifically sound criteria regarding where funds should most usefully be targeted are used to determine priorities for taxonomic research. It is therefore surprising that the processes of collecting, recognizing, and describing species are poorly understood and only rarely discussed (5–7) and that there is little research focused on the processes that result in the recognition of new species. Many groups of organisms are so poorly known that measuring any aspect of the discovery process suffers from lack of data. In terms of completing the species-level "inventory of life," the flowering plants are viewed as an attainable priority research target because they are already relatively well known and the final inventory is estimated to be only 10–20% from completion (8). Furthermore, plants are pivotal organisms for monitoring and measuring global biodiversity because they comprise a species-rich component of almost all habitats on earth (9). An enhanced scientific understanding of the discovery process for flowering plants could help define specific priorities for funding agencies and facilitate the meeting of global biodiversity targets. Here, we focus on the temporal dynamics of the lag between the collection of flowering plant specimens and their subsequent recognition and description as new species (7). For a representative dataset, the discovery time (I) between the date of the earliest specimen collected (C) and date the description was published (D) was calculated for each species (Fig. 1).

Results

Discovery I ranged from 1 to 210 y, averaging 38.8 y for monographs and 32.4 y for *Kew Bulletin*. Median I (the time taken to describe half the specimens collected in a particular year) was 22–25 y (95% confidence interval) for *Kew Bulletin* and 25–34 y for monographs (Fig. 2*A*). The combined data had a median I of 23–25 y, with only 14.4–16.9% (95% confidence interval) of species being described within 5 y of collection. This result emphasizes the relative importance of older collections for the discovery of new species of flowering plant.

The difference in the distribution of I between *Kew Bulletin* and monographs was statistically significant (Cox proportional hazards model; $P < 0.001$), with *Kew Bulletin* collection having a 9.6–33.8% (95% confidence interval) greater rate of description and, thus, smaller I, than the monographs. This difference was reflected in later C for the *Kew Bulletin* data (interquartile range 1938–1979 vs. 1931–1975 for monographs). Cox models showed that the description rate increased by 7.5–8.0% per year (95% confidence interval), i.e., more recently collected specimens had a greater chance of being described. Comparing modeled description rates for specimens collected in a given year (1956, the mean of C), the discovery process was similar for each source (Fig. 2*B*). Therefore, differences in the distribution of I between sources could be wholly accounted for by the fact that monographs contain some older specimens: Otherwise, the process of description operating in the two sources appears to be identical.

Our results imply that significant numbers of undescribed species have already been collected and are housed in herbaria, awaiting detection and description. Based on current estimates that ≈20% of species of flowering plant (≈70,000 species) remain undescribed (8), and with an approximate annual description rate of 2,000 species (10–12), all flowering plants should be described within 35 y, i.e., by 2045. Cox models showed a small but significant decrease in description probability with D between 1970 and 2010, by 0.5–1.1%·y^{-1} ($P < 0.000001$). Extrapolating this trend forward to 2045, the model predicts that a large fraction of those unknown species are already in the collections: 47–59% under a *Kew Bulletin* description rate, and 53–66% for a monographic approach.

Discussion

There are many reasons why older specimens representing new species remain undetected and undescribed in herbaria. In many cases, herbaria are overloaded and specimens are unprocessed and unavailable for study; expertise in particular taxa is often lacking, so new species are unnoticed, misplaced, or assigned to unidentified material at the end of each family. Some specimens are incomplete or lack flowers or fruits (7). In addition, specimens are sometimes identified as new species, annotated, and even given manuscript names but never described and published.

Author contributions: D.P.B., M.A.C., and R.W.S. designed research; M.A.C., J.R.I.W., A.H.W., D.J.H., G.T.P., T.D.P., N.K.B.R., and R.W.S. performed research; D.P.B., G.D., J.P., and R.W.S. analyzed data; and D.P.B., M.A.C., J.R.I.W., A.H.W., D.J.H., G.T.P., G.D., and R.W.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: robert.scotland@plants.ox.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1011841108/-/DCSupplemental.

Fig. 1. Herbarium specimen of *Strobilanthes frondosa* first collected (*C*) in 1924 from Burma (Cooper 5943A), published 70 y later (*D*) in 1994 (24). In this example, *I* equals 70 y. The specimen is from the Royal Botanic Garden Edinburgh (photo courtesy of Prashant Awale).

The data discussed here shows that a large number of newly described species were found among the older specimens in different herbaria (7). Many of these species only came to light when detailed comparison of the complete range of species in a particular clade took place during the course of monographic or revisionary studies. In some cases, it is the combination of more recent collections and hitherto unrecognized older collections that together provide the geographical and morphological evidence for a new species. Our data imply that, by necessity, the way to uncover new species in herbaria is through careful and ongoing examination of all specimens across the range of a taxon, as reflected in our sources (13–17). The Chrysobalanaceae monograph demonstrates how an intensive period of taxonomic activity has a significant effect on species discovery from a combination of old and recent collections (18–20). The first part of the monograph published in 1972 (15) described 90 new species, with an average difference in the age from first collection to description of 35 y. From 1972 to 1989, when an additional 6,795 specimens were studied that had mostly been collected since 1972, an additional 63 species were described, with an average difference in the age from first collection to description of 14 y. From 1990 to 2001, an additional 38 species were described from a further 4,996 mostly new specimens, with an average difference in the age from first collection to description of 10.5 y. This reduction in discovery (*I*) was due to the continuing presence of a taxonomic expert who could identify new species quickly within the context of an existing sound monographic treatment. In the absence of taxonomic expertise and a sound foundation taxonomic account as written in 1972, the reduction in discovery (*I*) would not have been possible. For many large groups of tropical plants, such taxonomic revisions have not been carried out, which means that recognizing new species from new collections is often not possible and, therefore, discovery (*I*) remains high.

Our results show that collecting and publishing descriptions of new species are two distinct parts of the discovery process that are largely dissociated. Only a small number of new species are recognized at the time of being collected, and these species are usually published within a relatively short period. However, the vast majority of new species are initially unrecognized and are subsequently described from herbarium specimens, often after a considerable lapse of time. This delay is because the description and delimitation of species is a comparative exercise and, therefore, new species can be reliably recognized only by reference to other closely related species after comparison with existing herbarium collections. This feature of the discovery process emphasizes the importance of channeling adequate funds to the

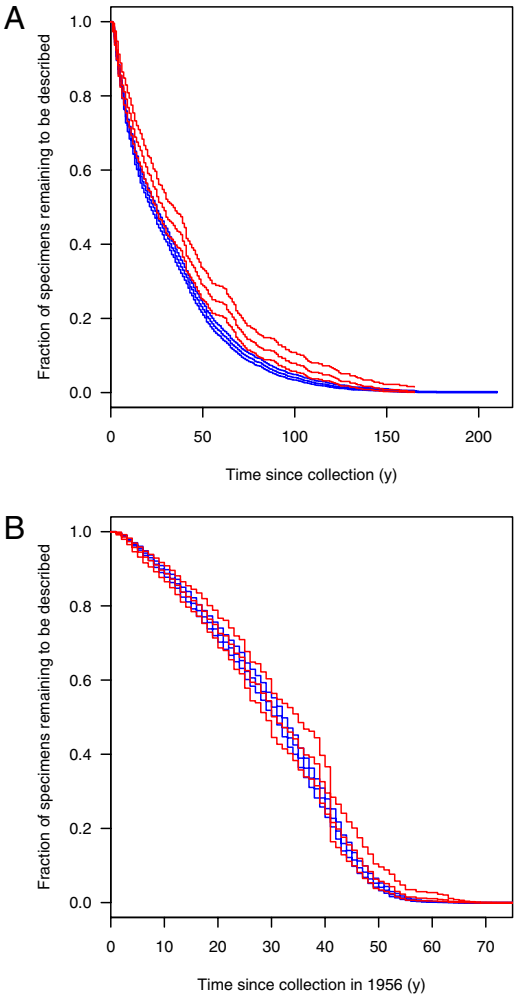

Fig. 2. Fraction of specimens remaining to be discovered against discovery time (*I*), for data from *Kew Bulletin* (blue lines) and Monographs (red lines). For each series, the central line shows the mean and the upper and lower lines show the 95% confidence limits on the mean. (*A*) All data: Mean *I* (and 95% confidence limits) to describe half the species is 23 y (22–25 y) for Kew and 28 y (25–34 y) for Monographs. (*B*) Specimens collected 1956 (the mean of *C*), fitted using Cox proportional hazards models: Mean *I* (and 95% confidence limits) to describe half the species is 32 y (31–33 y) for *Kew Bulletin* and 31 y (29–35 y) for Monographs. Note that the scales on the abscissa in *A* and *B* differ.

world's herbaria so that they can deal with the often substantial backlog of unprocessed collections while at the same time maintaining existing collections. Crucially our results highlight the central importance of taxonomic expertise that can sort, detect, and understand morphological variation in herbarium specimens.

To document fully the world's flora will require a combination of continuing field-work targeted at undercollected localities coupled with global taxonomic syntheses of major groups to discover and describe species that have escaped detection thus far. The absence of recent global, taxonomic accounts and expertise for many large tropical groups will be a major impediment for the completion of this task. In addition, herbaria may be reservoirs of undescribed diversity for relatively heavily collected floras (5–7). When the final plant collections have been made from the more inaccessible parts of the world, herbarium cabinets will still represent a final frontier for the discovery of a large number of new species of flowering plant. This fact emphasizes the pivotal role of herbarium-based taxonomic research activity in the documentation of the world's flora and the need

for widened access to global collections through the exchange and largescale digitisation (21) of existing specimens.

Materials and Methods

Data. Data were assembled for 3,219 species described during the period 1970–2010 and associated with specimens collected between 1770–2007 (*SI Text*). We chose this period because it most accurately reflects the contemporary situation and also avoids the complicated taxonomic history and synonymy associated with older species descriptions. The data were gathered from two sources that represent the full range of taxonomic activity and geography: new species (sp. nov.) from six monographic treatments (n = 449 species) and the journal *Kew Bulletin* (n = 2,770 species). We selected monographic treatments of taxa with a range of geographical distribution patterns to best capture global differences in species occurrence and the history of taxonomic activity, i.e., the pan-tropical Chrysobalanaceae, *Aframomum* from Africa, *Inga* from tropical America, *Strobilanthes* from South and South East Asia, *Agalmyla* from Malesia, and *Hypericum* distributed in temperate and subtropical regions, also extending into tropical montane habitats (Dataset S1). Five of the monographs included fieldwork and examination of large quantities of recently collected specimens. For example, after the first part of the Chrysobalanaceae monograph was published in 1972, 11,500 additional herbarium collections were made and then examined by the author. For *Aframomum,* 547 of 3,184 specimens examined were collected after 1990. For *Strobilanthes*, targeted field work was carried out in Sri Lanka, India, Bhutan, Java, and the Philippines over a 15-y period. We reasoned that new species described in *Kew Bulletin* provide a representative sample of all new species descriptions included in taxonomic revisions, small monographs, and novelties as a result of ongoing collecting activities. Any overlapping records from parts of monographs published in *Kew Bulletin* were identified and counted once only under monographs. The discovery time (I) between the date of the earliest specimen collected (C) and date the description was published (D) was calculated for each species.

Statistical Analysis. The process of discovery was investigated by using survival analysis, which examines and models the time it takes for events to occur (22, 23). Survival analysis is often applied to survival until death, but it can be applied to a wide range of situations in which individuals change state (for example, failure time of mechanical components). Because the data represent a change of state over time (from being a collected specimen in a herbarium to being a named species), the interval I can be analyzed by using these techniques. Survival curves (the fraction of specimens remaining to be named over time) and their variances were calculated by using the Kaplan–Meier estimator. Survival data can be modeled by using hazard functions, where the hazard h at time t is the instantaneous risk of state-change (in this case, description of a collected specimen), conditional on being collected by undescribed at that time:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr[(t \leq I < t + \Delta t) | I > t]}{\Delta t}.$$

Survival data can be modeled by using the log of the hazard function as the response variable and a linear function of log time as the predictor, which leads to the Weibull distribution of survival times:

$$\log h(t) = v + \rho \log(t).$$

Quantile plots indicate that I for the entire dataset and the collections separately match Weibull distributions closely, validating the use of survival analyses for these data (Fig. S1).

A common method to analyze the effect of covariates on the hazard function is through Cox proportional hazards models, where the baseline hazard function, log $h_0(t)$, is modified by covariates:

$$\log h_i(t) = \log h_0(t) + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}.$$

Here, β are coefficients, i is a subscript for observation, and x is a covariate. Cox models are therefore a form of General Linear Model. The effect of C and D on the hazard function and predictions of future (I) were estimated by using Cox models.

ACKNOWLEDGMENTS. Fred Barry, Kevin Gaston, Nicholas Harberd, Stephen Harris, Jane Langdale, and two anonymous reviewers provided useful comments for earlier versions of this paper. We thank Anne Sing and Denis Filer for help with data entry and manipulation. R.W.S. acknowledges the Royal Society for funding during the early period of the *Strobilanthes* monograph in the form of a University Research Fellowship.

1. Chapman AD (2009) Numbers of living species in Australia and the world. *Australian Biological Resources Study* (Canberra, Australia).
2. Pitman NCA, Jørgensen PM (2002) Estimating the size of the world's threatened flora. *Science* 298:989.
3. Stuart SN, Wilson EO, McNeely JA, Mittermeier RA, Rodríguez JP (2010) Ecology. The barometer of life. *Science* 328:177.
4. House of Lords (2008) *Systematics and Taxonomy* (The Stationery Office, London).
5. Ertter B (2000) Floristic surprises in North America north of Mexico. *Ann Mo Bot Gard* 87:81–109.
6. Hartman RL, Nelson BE (1998) Taxonomic novelties from North America north of Mexico: A 20-year vascular plant diversity baseline. *Monographs in Systematic Botany from the Missouri Botanic Garden* 67:1–59.
7. Shevock J, Taylor DW (1987) Conservation and management of rare and endangered plants. *Proceedings of a California Conference on the Conservation and Management of Rare and Endangered Plants*, ed Elias TS (Calif Native Plant Soc, Sacramento, CA), pp 91–98.
8. Joppa LN, Roberts DL, Pimm SL (July 7, 2010) How many species of flowering plants are there? *Proc Biol Sci*, 10.1098/rspb.2010.1004.
9. Mutke J, Barthlott W (2005) Patterns of vascular plant diversity at continental to global scales. *Biologiske Skrifter* 55:521–531.
10. International Institute for Species Exploration (2008) 2008 SOS State of Observed Species: A report card on our knowledge of earth's species. Available at http://species.asu.edu/SOS_2008. Accessed February 8, 2010.
11. Paton AJ, et al. (2008) Towards target 1 of the global strategy for plant conservation: A working list of all known plant species - progress and prospects. *Taxon* 57:602–611.
12. Prance GT (2001) Discovering the plant world. *Taxon* 50:345–359.
13. Hilliard OM, Burtt BL (2002) The genus *Agalmyla* (Gesneriaceae-Cyrtandroideae). *Edinb J Bot* 59:1–210.
14. Pennington TD (1997) *The Genus Inga* (R Botanic Garden, Kew, UK).
15. Prance GT (1972) Chrysobalanaceae. *Flora Neotropica* 9:1–410.
16. Robson NKB (1990) Studies in the genus *Hypericum* L. (Guttiferae) 8. Sections 29. Brathys (part 2) and 30. Trigynobrathys. *Bull Nat Hist Mus Bot* 20:1–151.
17. Wood JRI, Scotland RW (2009) New and little known species of *Strobilanthes* (Acanthaceae) from India and South East Asia. *Kew Bull* 64:3–47.
18. Prance GT (1977) Floristic inventory of the tropics: Where do we stand? *Ann Mo Bot Gard* 64:659–684.
19. Prance GT (1984) *Current Concepts in Plant Taxonomy*, eds Heywood VH, Moore DM (Academic, London), pp 365–397.
20. Prance GT, Cambell DG (1988) The present state of tropical floristics. *Taxon* 37: 519–548.
21. Wheeler QD (2008) *The New Taxonomy*, ed Wheeler QD, Systematics Association Special Volume 76, (CRC, Boca Raton, FL), pp 211–226.
22. R Development Core Team (2010) http://www.r-project.org. Accessed January 7, 2010.
23. Therneau T (2009) http://cran.r-project.org/web/packages/survival/index.html. Accessed January 7, 2010.
24. Wood I, Jr (1994) Notes relating to the flora of Bhutan: XXIX. Acanthaceae, with special reference to Strobilanthes. *Edinb J Bot* 51:175–274.