Megataxa 001 (1): 035–038
https://www.mapress.com/j/mt/

Correspondence

ISSN 2703-3082 (print edition)
MEGATAXA
ISSN 2703-3090 (online edition)

https://doi.org/10.11646/megataxa.1.1.6

The promise of next-generation taxonomy

MIGUEL VENCES
Zoological Institute, Technische Universität Braunschweig, Mendelssohnstr. 4, 38106 Braunschweig, Germany
m.vences@tu-braunschweig.de; https://orcid.org/0000-0003-0747-0817

Documenting, naming and classifying the diversity of life on Earth provides baseline information on the biosphere, which is crucially important to understand and mitigate the global changes of the Anthropocene. Since Linnaeus, taxonomists have named about 1.8 million species (Roskov *et al.* 2019) and continue doing so at a rate of about 15,000–20,000 species per year (IISE 2011). Natural history collections—museums, herbaria, culture collections and others—hold billions of collection specimens (Brooke 2000) and have teamed up to assemble a cybertaxonomic infrastructure that mobilizes metadata and images of voucher specimens, now even at the scale of digitizing entire collections of millions of insect or herbaria vouchers in automated imaging lines (e.g., Tegelberg *et al.* 2014; Heerlin *et al*. 2015). Hundreds of millions of specimen metadata are available from data aggregators such as GBIF and iDigBio (Nelson and Ellis 2018), and much progress has especially been achieved with building curated species databases (Crous *et al.* 2004; Patterson *et al*. 2010), dozens of which for instance contribute to the Catalogue of Life (Roskov *et al.* 2019). These are impressive achievements and taxonomy certainly qualifies as big data science by fulfilling the main criteria of volume, variety, and velocity (De Mauro *et al.* 2016).

But are we doing enough? Barely so. Given the unknown and disputed, yet undoubtedly enormous proportion of undocumented and unnamed species (Larsen *et al.* 2017), taxonomy at its current pace will not be able to deliver in a reasonable time frame a fairly complete, or at least representative, inventory of species on the Globe—which I am convinced is needed to reliably inform the assessment of biodiversity patterns, anthropocenic changes of species composition, declines, and priority regions for conservation. Wheeler *et al.* (2012) defined an ambitious goal to name 10 million species in less than 50 years. This translates into naming 200,000 species per year, one order of magnitude more than the current rate.

Can this be achieved? Although many will skeptically shake their head I believe yes, but it requires new thinking and renovation of some established work procedures and concepts. We should meet three main challenges, using new technological developments without throwing the well-tried and successful foundations of Linnaean nomenclature overboard.

1. Fully embrace cybertaxonomy, machine learning and DNA taxonomy to ease, not burden the workflow of taxonomists.

Computer power and especially, DNA sequencing capacity increases faster than exponentially (e.g., Rupp 2018) and new technologies offer unprecedented opportunities for classifying specimens based on molecular evidence or image analysis. Yet, the vast majority of species are still named without molecular evidence, and most original data produced in the alpha taxonomic workflow, especially images and measurements, are not submitted to repositories for future re-use (A. Miralles, M. Vences and collaborators, unpublished data). Cybertaxonomy is flourishing and provides many benefits to end-users of taxonomy, but so far is apparently only of limited utility for taxonomists themselves. Maybe we are not recognizing the opportunities that the new digital and high-throughput methods are offering?

For many taxonomists, using data-rich -omics techniques in their usual workflow, primarily signifies additional burden, probably explaining their reluctance to employ such approaches. In 2018, an average taxonomic paper in *Zootaxa* named about 3 new species and had 3 authors (A. Miralles, M. Vences and collaborators, unpublished data). For such small teams and projects, submitting sequences to GenBank, linking specimens to unique stable identifiers (Güntsch *et al*. 2018), uploading images and microCT scans to repositories is a highly inefficient use of time and resources. Working in large teams, with colleagues specializing in each of these tasks, and routinely applying them to massive fast-track taxonomic projects of naming hundreds if not thousands of species at once (e.g., Riedel *et al*. 2013) could be the key. We should learn from particle physicists or genome researchers who have long understood the prospect of

Submitted: 16 Dec. 2019; accepted by Z.-Q. Zhang: 20 Jan. 2020; published: 31 Jan. 2020

tackling big science in big consortia of authors. *Megataxa* is a huge step into the right direction, and the multi-author papers jointly naming multiple new species of fungi, e.g., in Fungal Planet / Persoonia (http://www.fungalplanet.org/), Fungal Diversity (Fungal Diversity Notes) and Phytotaxa, have paved the way.

Some taxonomists have claimed that "the description of the entire world's species is quite simply impossible" (Bickel 2009) because "taxonomic description has traditionally been haphazard, based on the interests and passions of individual taxonomists". Probably, indeed, naming all extant species is impossible—especially because many local endemics will go extinct before we get the chance to collect them—but we certainly can obtain a much more complete and representative inventory than we have now. Of course, next-generation taxonomy must retain niches for passion-driven taxonomists who, motivated purely by curiosity, revise in great detail the diversity and morphology of specific groups. But in other taxa, including many "orphans" without active specialists studying them, only industrial-scale inventories can reveal their true richness and connected to this, their roles and services to ecosystems. Scientific curiosity and enthusiasm can also be fueled by discovering such patterns, and by the sheer astonishment on the enormous species numbers that will be unveiled. Maybe this perspective could even pull a very different, new group of biologists towards taxonomy?

2. Emphasize diagnosis over description, images over words.

In the jargon of taxonomists, we use the term "species description" for the process of discovering, diagnosing and naming a species in agreement with the rules of the *Codes*. This reflects that new species names today are accompanied by verbose morphological descriptions rather than simple diagnoses as initially used by Linnaeus. I posit we should re-consider this terminology and tradition, and herein have avoided the term "species description" which is so deeply embedded in our mind. As convincingly advocated by Renner *et al*. (2016), the emphasis should be less on description and more on diagnosis—high-resolution images could replace most of our traditional description of characters and thereby speed up the work without compromising precision. Often a high-resolution image coupled to a DNA barcode is more informative than broken fragments of a poorly preserved holotype.

Images could also be at the core of tackling the "dark diversity" of many understudied groups of taxa—if we make use of the full latitude offered by the rules in the Codes. The importance of voucher specimens and culture collections is beyond doubt, but there could be cases where it is worth considering the use of images to replace non-preserved type specimens. Leaving aside the odd threatened primate or bird for which no collecting permit can be obtained, this may apply in particular to protists or microscopic metazoans where non-destructive genetic sampling is almost impossible. Here, high-resolution imaging followed by single-cell transcriptomics or genomics will provide a rich set of data that may well justify naming a new species, even without depositing a cell culture or slide with a preserved type. We could learn a lot about such species: besides knowing at least some of their genes, we could delimit their geographical distribution, habitat and ecological interactions by sequencing environmental DNA using metabarcoding or metagenomic approaches (e.g., Srivathsan *et al.* 2016), and link all this to a traditional Linnean name.

3. Understand promises and pitfalls of omics approaches to avoid taxonomic inflation.

High throughput approaches have revolutionized many fields of biology and offer unprecedented opportunities for taxonomists. We are moving towards affordable DNA barcoding of millions of individuals, and the uncritical use of simplistic species delimitation methods e.g. based on "barcoding gaps" to such data sets is very tempting. It also is becoming possible to routinely sequence thousands of markers or full genomes which allows for ever more fine-scale analysis of population structure—not to be confused with species (Sukumaran & Knowles 2017). Speeding up the process of naming species must not compromise the quality of species hypotheses, and requesting a clear justification of species status would be a first, important step to deter taxonomic inflation—why are the characters in the diagnosis relevant enough to consider a lineage as distinct species under a certain species criterion (de Queiroz 2007)? To avoid excesses and initial rejection of the approach by the community of taxonomists, a conservative Biological Species Criterion could be implemented in high-throughput taxonomic pipelines—*i.e.*, naming only lineages that we can be confident are reproductively isolated. Reproductive isolation could be estimated by demanding, besides coalescent delimitation, genomic divergences above a high, conservative threshold (e.g. higher than average divergence among well-established species); or in other cases by sympatric occurrence without genomic admixture (Padial *et al*. 2010).

In addition, increased efforts should be directed to develop user-friendly bioinformatic tools for taxonomy. In comparison to the breathtaking pace at which new programs are published in phylogenetics, the selection

of software tailored to facilitate the alpha-taxonomic workflow is extremely limited. Many modern species delimitation approaches require using a complex pipeline of various programs, and none of them truly mirrors the workflow followed by integrative taxonomists (Padial *et al*. 2010) who for instance consider evidence from sympatry of non-admixing lineages as conclusive evidence of species distinctness, and evaluate the taxonomic relevance of morphological characters based on their variation in well-known species or on their relevance for mate choice. Software packages (a) combining geographical and morphological evidence with molecular species delimitation, (b) based on machine learning but also allowing users to input evidence from a diverse set of data, and (c) including machine-accessible portals for archiving and retrieving taxonomic data aggregated around specimen identifiers (extended specimens or cyberspecimens; Lendemer *et al*. 2019), would have an enormous potential to improve species delimitation at the large scales required to speed up the naming of Earth's species diversity.

There is little doubt that large-scale assessments of "dark diversity" or "open-ended taxa" will increasingly be carried out using molecular tools, and provide fascinating insights into diversity patterns *e.g.* among habitat types or biogeographic regions (*e.g.*, Srivathsan *et al.* 2019). We can choose: Are we satisfied with such analyses resulting in, and based on, poorly defined molecular operational taxonomic units (mOTUs) or DNA barcode index numbers (BINs) (Ratnasingham & Hebert 2013)? My plea is to take just one additional step, add high-quality images, demand high-quality species delimitation, and name these units under a Linnaean scheme. As claimed by Bik (2017), if we play our cards right, taxonomy could be on the brink of another golden age. For this to work out, we taxonomists must take matters into our own hands—and define more precisely which tools we need for improving speed and quality of alpha taxonomy.

Acknowlegments

I am grateful to numerous colleagues, in particular Aurélien Miralles, Alexander Riedel, Susanne S. Renner, and Mark D. Scherz for stimulating discussions during the past years, and to Daniel J. Bickel for his critical but constructive review of a previous draft of the manuscript. This work benefited from the sharing of expertise within the priority program SPP 1991 "Taxon-Omics" of the Deutsche Forschungsgemeinschaft.

References

Brooke, M. de L. (2000) Why museums matter. *Trends in Ecology and Evolution,* 15, 136–137. https://doi.org/10.1016/S0169-5347(99)01802-9

Bickel, D. (2009) Why *Hilara* is not amusing: the problem of open-ended taxa and limits of taxonomic knowledge. *In*: Pape, T., Bickel, D. & Meier, R. (eds.) *Diptera Diversity: Status, Challenges, and Tools.* Brill, Leiden, pp. 279–301.

Bik, H.M. (2017) Let's rise up to unite taxonomy and technology. *PLoS Biology,* 15, e2002231. https://doi.org/10.1371/journal.pbio.2002231

Crous, P.W., Gams, W., Stalpers, J.A., Robert, V. & Stegehuis, G. (2004) MycoBank: an online initiative to launch mycology into the 21st century. *Studies in Mycology,* 50, 19–22.

de Queiroz, K. (2007) Species concepts and species delimitation. *Systematic Biology,* 56, 879–886. https://doi.org/10.1080/10635150701701083

De Mauro, A., Greco, M. & Grimaldi, M. (2016) A formal definition of Big Data based on its essential features. *Library Review,* 65, 122–135. https://doi.org/10.1108/LR-06-2015-0061

Güntsch, A., Groom, Q., Hyam, R., Chagnoux, S., Röpert, D., Berendsohn, W., Casino, A., Droege, G., Gerritsen, W., Holetschek, J., Marhold, K., Mergen, P., Rainer, H., Smith, V. & Triebel, D. (2018) Standardised globally unique specimen identifiers. *Biodiversity Information Standards,* 2, e26658. https://doi.org/10.3897/biss.2.26658

Heerlien, M., Van Leusen, J., Schnörr, S., De Jong-Kole S, Raes, N. & Van Hulsen, K. (2015) The natural history production line: An industrial approach to the digitization of scientific collections. *ACM Journal on Computing and Cultural Heritage,* 8, 3. https://doi.org/10.1145/2644822

IISE (2011) State of Observed Species. Tempe, AZ. International Institute for Species Exploration. Available from: http:/species.asu.edu/SOS (Accessed 15 March 2019).

Larsen, B.B., Miller, E.C., Rhodes, M.K. & Wiens, J.J. (2017) Inordinate fondness multiplied and redistributed: the number of species on Earth and the new pie of life. *Quarterly Review of Biology,* 92, 229–265. https://doi.org/10.1086/693564

Lendemer, J., Thiers, B., Monfils, A.K., Zaspel, J., Ellwood, E.R., Bentley, A., LeVan, K., Bates, J., Jennings, D., Contreras, D., Lagomarsino, L., Mabee, P., Ford, L.S., Guralnick, R., Gropp, R.E., Revelez, M., Cobb, N., Seltmann, K. & Aime, M.C. (2019) The extended specimen network: a strategy to enhance US biodiversity collections, promote research and education. *BioScience*, biz140, https://doi.org/10.1093/biosci/biz140

Nelson, G. & Ellis, S. (2018) The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B,* 374: 20170391. https://doi.org/10.1098/rstb.2017.0391

Padial, J.M., Miralles, A., De la Riva, I. & Vences, M. (2010) The integrative future of taxonomy. *Frontiers in Zoology,* 7, e16. https://doi.org/10.1186/1742-9994-7-16

Patterson, D.J., Cooper, J., Kirk, P.M., Pyle, R.L. & Remsen, D.P. (2010) Names are key to the big new biology. *Trends in Ecology and Evolution,* 25, 686–691. https://doi.org/10.1016/j.tree.2010.09.004

Renner, S.S. (2016) A return to Linnaeus's focus on diagnosis, not

description: The use of DNA characters in the formal naming of species. *Systematic Biology,* 65, 1085–1095. https://doi.org/10.1093/sysbio/syw032

Riedel, A., Sagata, K., Surbakti, S., Tänzler, R. & Balke, M. (2013) One hundred and one new species of *Trigonopterus* weevils from New Guinea. *Zookeys*, 280, 1–150. https://doi.org/10.3897/zookeys.280.3906

Roskov, Y., Ower, G., Orrell, T., Nicolson, D., Bailly, N., Kirk, P.M., Bourgoin, T., DeWalt, R.E., Decock, W., Nieukerken, E. van, Zarucchi, J. & Penev, L. (Eds) (2019) Species 2000 & ITIS Catalogue of Life, 26th February 2019. Digital resource at www.catalogueoflife.org/col. Species 2000: Naturalis, Leiden, the Netherlands. ISSN 2405-8858.

Rupp, K. (2018) 42 Years of Microprocessor Trend Data. Website. Available from: https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/ (Accessed 13 March 2019)

Ratnasingham, S., Hebert, P.D. (2013) A DNA-based registry for all animal species: the barcode index number (BIN) system. *PLoS One*, 8, e66213. https://doi.org/10.1371/journal.pone.0066213

Srivathsan,A., Ang, A., Vogler, A.P. & Meier, R. (2016) Fecal metagenomics for the simultaneous assessment of diet, parasites, and population genetics of an understudied primate. *Frontiers in Zoology,* 13, 17. https://doi.org/10.1186/s12983-016-0150-4

Srivathsan, A., Hartop, E., Puniamoorthy, J., Lee, W.T., Kutty, S. N., Kurina, O. & Meier, R. (2019) Rapid, large-scale species discovery in hyperdiverse taxa using 1D MinION sequencing. *BMC Biology*, 17, 96. https://doi.org/10.1186/s12915-019-0706-9

Sukumaran, J. & Knowles, L.L. (2017) Multispecies coalescent delimits structure, not species. *Proceedings of the National Academy of the United States of America*, 114, 1607–1612. https://doi.org/10.1073/pnas.1607921114

Tegelberg, R., Mononen, T. & Saarenmaa, H. (2014) High-performance digitization of natural history collections: Automated imaging lines for herbarium and insect specimens. *Taxon*, 63, 1307–1313. https://doi.org/10.12705/636.13

Wheeler, Q.D., Knapp, S., Stevenson, D.W., Stevenson, J., Blum, S.D. , Boom, B.M., Borisy, G.G., Buizer, J.L., De Carvalho, M.R., Cibrian, A., Donoghue, M.J., Doyle, V., Gerson, E.M., Graham, C.H., Graves, P., Graves, S.J., Guralnick, R.P., Hamilton, A.L., Hanken, J., Law, W., Lipscomb, D.L., Lovejoy, T.E., Miller, H., Miller, J.S., Naeem, S., Novacek, M.J., Page, L.M., Platnick, N.I., Porter-Morgan, H., Raven, P.H., Solis, M.A., Valdecasas, A.G., Van Der Leeuw, S., Vasco, A., Vermeulen, N., Vogel, J., Walls, R.L., Wilson, E.O. & Woolley, J.B. (2012) Mapping the biosphere: exploring species to understand the origin, organization and sustainability of biodiversity. *Systematics and Biodiversity,* 10, 1–20. https://doi.org/10.1080/14772000.2012.665095