How Much of Biodiversity is Represented in Collections: A big data workflow of aggregated occurrence data

Pieter Huybrechts‡, Maarten Trekels‡, Quentin Groom‡

‡ Meise Botanic Garden, Meise, Belgium

Corresponding author: Pieter Huybrechts (pieter.huybrechts@plantentuinmeise.be)

Abstract

Natural history collections play a pivotal role in taxonomy, which in turn supports all of biology, but particularly conservation and biodiversity policy. However, to provide this role, it is necessary to know what specimens are stored where, and how complete the collection is. The biodiversity held within collections globally remains uncertain, with an estimated 1.2 to 2.1 billion (10^9) specimens (Ariño 2010), of which around 200 million are represented on the Global Biodiversity Information Facility (GBIF)**.** Here we estimate the total biodiversity in collections worldwide by extrapolating from those specimens we know about.

Data aggregators such as GBIF provide an ever-changing window into the contents of collections. We use non-parametric estimators that allow for the approximation of the number of classes in an incomplete set, such as the number of species within a collection, but also the proportion of biodiversity preserved on a national or continental level (for example within a taxonomic group, compared to the world or to a continent). Because the contents of data aggregators such as GBIF, are in constant flux, our workflow is made to be repeatable on the monthly snapshots that GBIF provides.

The results of the workflow expose data gaps in GBIF, namely that collections from some large geographical regions, such as Asia, are poorly represented, but also taxonomic gaps exist, such as several Coleoptera families where many more species are accepted in the backbone than are represented on GBIF. As more data are published to GBIF the estimates for these taxon groups and geographical regions will improve. The detection of data gaps within data aggregators such as GBIF, and the subsequent mobilisation of missing information remains a priority for both aggregators and researchers (GBIF Secretariat 2022, Collen et al. 2008, Hochkirch et al. 2020). Our workflow will allow for continuous monitoring of collections and groups of collections of their coverage of global biodiversity, and the results can inform their collection development strategy.

Keywords

natural history collections, GBIF, extrapolation, data gap

Presenting author

Pieter Huybrechts

Presented at

TDWG 2022

Funding program

This work was facilitated by the Research Foundation – Flanders (FWO) as part of the Flemish contribution to the DiSSCo Research Infrastructure under grant n° I001721N

Conflicts of interest

References

- Ariño A (2010) Approaches to estimating the universe of natural history collections data. Biodiversity Informatics 7 (2). https://doi.org/10.17161/bi.v7i2.3991
- Collen B, Ram M, Zamin T, McRae L (2008) The Tropical Biodiversity Data Gap: Addressing Disparity in Global Monitoring. Tropical Conservation Science 1 (2): 75-88. https://doi.org/https://doi.org/10.1177%2F194008290800100202
- GBIF Secretariat (2022) GBIF Work Programme 2022: Annual Update to Implementation Plan 2017–2022. GBIF https://doi.org/10.35035/doc-jjrz-b144
- Hochkirch A, Samways M, Gerlach J, Böhm M, Williams P, Cardoso P, Cumberlidge N, Stephenson PJ, Seddon M, Clausnitzer V, Borges PV, Mueller G, Pearce-Kelly P, Raimondo D, Danielczak A, Dijkstra K (2020) A strategy for the next decade to address data deficiency in neglected biodiversity. Conservation Biology 35 (2): 502-509. https://doi.org/10.1111/cobi.13589