How to do meta-analysis of open datasets

The amount of open data in ecology and evolution is increasing rapidly, yet this resource remains underused. Here, we introduce a new framework and case study for conducting meta-analyses of open datasets, and discuss its benefits and current limitations.

Antica Culina, Thomas W. Crowther, Jip J. C. Ramakers, Phillip Gienapp and Marcel E. Visser

In recent decades, the meta-analysis approach has emerged as the most valuable avenue for scientific progress, along with empirical studies and theoretical models[1,2]. Traditional meta-analysis combines results from a number of studies (ideally all) conducted on the same research question, to statistically summarize findings, evaluate discrepancies and detect generalizable effects[2]. The ability to detect overarching patterns makes meta-analyses extremely relevant to evolutionary ecology, which is characterized by highly complex systems, heterogeneous environments and variable methodologies[3,4].

Systematic advances in the meta-analysis approach over the past decade have been intended to improve the transparency, replicability, reliability and impact of data synthesis efforts[2,5–7]. However, despite these advances, the major outstanding limitation of any synthesis remains the challenge of accessing a comprehensive range of available data on the topic[7]. Conventionally, meta-analyses are conducted using effect sizes (that is, measure of the strength and direction of effects) extracted from the values reported in published studies. These meta-analyses are often limited to studies that focus specifically on the topic of interest (we term these 'target studies'). However, a wealth of useful data is often available in various 'non-target studies' that have attained relevant information to address different research questions. Additional data from non-target studies can enhance the statistical power of meta-analyses (a fact that has been widely accepted and embraced in medical research[8]), as well as considerably reduce current issues with biased effect sizes. These data can be used either on their own, or in a combination with data from target studies. Until now, the complex and variable research landscape in ecology and evolution has restricted such data ingestion from non-target studies. However, the increase in data made openly accessible, as now required by many journals, is transforming our capacity to access, evaluate and use raw data from both target and non-target studies. Hence, our potential to survey the data landscape and gain a comprehensive understanding of the available information has never been greater[9]. Yet, unlike other scientific fields, this resource remains relatively unexploited in the field of ecology and evolution[10,11].

Data retrieval for meta-analysis

Here, we describe how to transparently retrieve and select data, when the information retrieval starts from published (open) datasets, rather than from published studies. Our standard is based on existing guidelines for the information retrieval in ecological/evolutionary meta-analysis[5,6,12,13], but adapted specifically for open data. The retrieval and selection process should be highly transparent — we provide a checklist of the information that needs to be recorded (Table 1). This information should ideally be supported by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses[14] diagram (Supplementary Fig. 1).

In the first step of the approach (Step 1), researchers need to identify the type of data needed to answer the meta-analysis question (or test hypothesis), set appropriate exclusion/inclusion criteria and choose the search terms (used in a search for the relevant data). This is followed by the data search. In evolutionary ecology, datasets are usually scattered across various repositories (for example, Dryad, Figshare, Zenodo) or published in the supplementary materials associated with a paper. Thus, an effective search should be conducted using data-harvesting platforms that crawl through many different research data repositories that host research data (like Web of Science crawls through journals in a search for articles); some also explore supplementary materials of published papers for additional information. A complete overview of how to navigate the data landscape by using data search platforms can be found in ref. [15]. We suggest using DataCite, BASE search engine and DataONE (see Box 1). The original search terms usually need to be adjusted according to the output of the initial

Table 1 | Checklist of the main steps in conducting meta-analysis that starts from datasets

Step	What to record (report)
Step 1: what type of data are needed and where/how to obtain them?	Research question/questions The exact exclusion/inclusion criteria Platform(s) used in search Search terms and syntax (for every platform; whether and how search terms were adjusted)
Step 2: screening the results according to the meta-data provided (keywords, dataset title, description of the dataset and/or subject area)	What meta-data screening was based on Number of excluded results Reasons for exclusion (optional)
Step 3: open and screen remaining datasets	Number of excluded results Reasons for exclusion (optional)
Step 4: detailed examination of the datasets. Contacting the authors of the dataset about missing/unclear information	Number of excluded results Reasons for exclusion Whether the authors were contacted and with what outcome
Step 5: calculate the effect sizes	Statistical procedures to calculate effect sizes
Step 6: contact the authors to check if they agree with the approach	Contact letter, author responses, dates of contact Datasets excluded based on authors' feedback and reasons why
Step 7: conduct the statistical part of meta-analysis	The dataset used in meta-analysis Exact models/formulas

Following these steps will ensure a transparently conducted meta-analysis that complies with the current scientific standards.

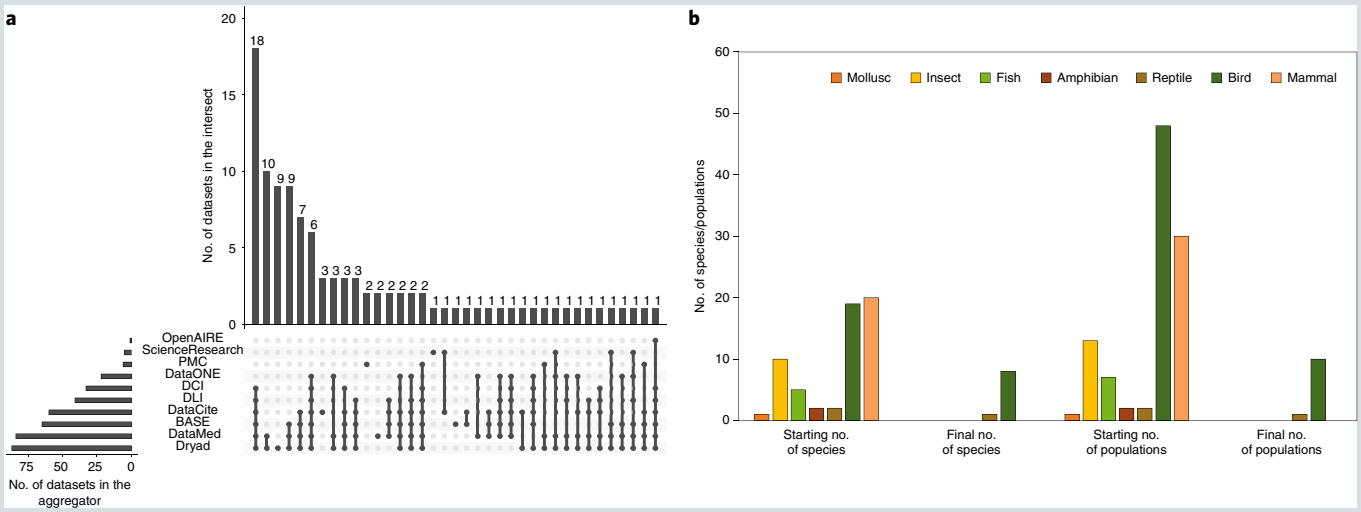
Box 1 | Application of the framework: environment, heritability and selection

While the environment has been shown to influence both selection and heritability of animals (see review in ref. [22]), the number of studies exploring both in the same systems is limited[23,24]. As such, there were not enough published studies on this topic to synthesize and generalize the relationship between selection and heritability[22]. To address this question, we[17] needed data on (1) pedigrees or additive genetic relatedness of these individuals, (2) individual phenotypes, (3) individual fitness and (4) the environment (which we defined by averaging the phenotypes across the population in a given year, and variance-standardizing it across years). We expected that pedigree data would represent the limiting source of data. Therefore, we started our search for this type of data deposited in online databases. The details on the data search and data screening process are provided in ref. [17] and its supplementary material. Here, we provide a summary of the search and data collection process.

After searching through 12 different aggregators of research data repositories (Europe PMC, DataCite, BASE, OpenAIRE, ScienceResearch, DataONE Mercury search, Web of Science Data Citation Index (DCI), SciELO, Research Data Australia, DLI service, Dryad Digital Repository and DataMed) and screening through the original search results (steps 1–3), we located 103 animal pedigreed datasets. Different aggregators we consulted identified different parts of the overall sample of datasets (panel a of the figure).

However, after a careful examination of these original 103 pedigree datasets, we were forced to discard 88 because of (1) embargoed data, corrupted or 'encrypted' files, (2) insufficient number of individuals with pedigree or (3) phenotypic information, (4) lack of natural environmental variation in the phenotype (this excluded all laboratory populations), (5) too few years included in the dataset (at least six years) or (6) other issues (for example, non-matching identifications of animals in pedigree and phenotype file) (see Table S1 of the supplementary material in ref. [17]). This left us with 15 datasets for analysis. After analysis, we contacted the original data owners to check whether we had misinterpreted their data, as this was one of the main concerns about the use of open data[16]. Based on the authors' feedback, we excluded another four datasets for various reasons related to non-random exclusion of individuals from the dataset (potentially leading to biased fitness measurements and quantifications of the environment). The reduction in the overall sample size, from 103 pedigree results obtained to 11 that we could use in the analysis, drastically reduced the number of taxa and populations represented in the dataset (amphibians, fish, insects, mammals and molluscs; panel b of the figure).

We conducted an additional literature search to identify studies that potentially also contained pedigree data (see ref. [17]), identifying three additional datasets yielding data on 49 traits in 15 populations of nine species in total. Overall, we e-mailed owners of 18 datasets to check if they agreed with the way we analysed their data. The majority (16) of them offered advice on the analysis (also leading to the exclusion of four datasets, see above) and were supportive towards our efforts, while two were negative towards the use of their data. After conducting the analysis (including heterogeneity analysis), we found that any effects of the environmentally caused coupling between heritability and selection on expected evolutionary response were small.



Summary of the results of the pedigree dataset search and dataset loss in the selection process. a, Number of pedigree datasets found by each aggregator (left-hand side frequency diagram) and the number of datasets in intersections among aggregators (main frequency diagram). For example, the same 18 datasets were obtained by DCI, DLI, DataCite, BASE, DataMed and Dryad search (first horizontal bar), while the Dryad search resulted in nine unique datasets. **b**, Loss of species and populations of different taxonomical groups (molluscs, insects, fish, amphibians, reptiles, birds and mammals) from the original 103 pedigree datasets obtained by search for open pedigree data to the final 11 datasets used in the analysis.

search (for example, when some obviously irrelevant terms appear in the search results). After the initial search, duplicates can be eliminated.

The next steps (2–4) describe the screening of the obtained datasets. This starts with screening according to the meta-data (data that describe the dataset) provided by the search platform (Step 2); these will vary between different search platforms (usually keywords, dataset title, dataset description and/or subject area).

Thus, it is important to record and report on which meta-data the screening was based. This step is equivalent to the initial screening of the title, abstract and keywords in the 'traditional' meta-analysis that starts from published studies. The main difference is that the standards to describe datasets are less well established than the standards to describe articles (title, abstract, keywords, subject areas). Thus, this screening might be more time-consuming, and lead to the retention of more irrelevant datasets. Next (Step 3), each potentially relevant dataset should be opened and screened to identify whether it corresponds to meta-analysis requirements.

The remaining datasets are relevant according to the dataset type, but some will be excluded (Step 4) as they do not match the specific inclusion criteria or are not fit for use because information crucial to run the desired analysis (to obtain the effect sizes) is missing (equivalent to under-reported effects in the approach that starts from published studies). At this stage, researchers might decide to contact the dataset owner(s).

The final list of datasets is then used to calculate the effect sizes (Step 5). Ideally, all effect sizes are calculated in the same, standardized way. This process can take several sub-steps. In line with good scientific practice, and to address the issues of data misinterpretation[16], owners of the datasets should be contacted, at the latest, when analysing their data, and asked whether they agree with the way in which the data were processed (Step 6). Some data owners specifically ask (in the meta-data files) to be contacted directly if there are plans to use the data. Some datasets might be excluded after this step. Statistical analyses can then be conducted using these effect sizes (Step 7) following the existing guidelines (choose an appropriate model, explore the sources of heterogeneity, account for non-independencies and, if considered necessary, test for publication bias[12,13]). Statistical analysis can also be conducted using both, effect sizes calculated from raw data and those calculated using values reported in published articles (when possible). In this case, information retrieval protocol should be recorded separately for the data and article selection process[6,12,13], and we would further advise controlling for the source of effect size (data or article) when conducting the statistical analysis.

To demonstrate the information retrieval framework, in Box 1 we outline the search for pedigree datasets for the meta-analysis that aimed at evaluating the strength of the evidence for the environmental coupling of heritability and selection[17].

Benefits of open data meta-analysis

Our case study (Box 1) demonstrates an obvious benefit of the information retrieval that starts from published data (rather than published studies): the considerable increase in the data available to conduct meta-analysis (and thus in the number of research questions that can be addressed[18]). These data can be used on their own to calculate effect sizes for the meta-analysis, or used alongside effect sizes extracted from published studies. In our example, a traditional meta-analysis was impossible (only two published studies on the research question, see Box 1). Use of open data from studies that themselves addressed another question enabled us to collect enough evidence for meta-analysis. Given that the number of published datasets is greatly increasing across evolutionary and ecological fields[9,15], the scope of evolutionary ecology meta-analysis can be extended, and not limited only to target studies in the published literature.

An additional benefit of open data is the reduced publication bias that stems from the selective reporting of 'significant' or 'interesting' results[7]. The under-reporting of weak, negative or unwanted effects (or ambiguous results) is common across scientific disciplines: two reviews showed that basic information (sample size and variance) was missing from generally half of otherwise relevant primary studies collected for meta-analysis in conservation ecology[19] and evolutionary ecology[20]. Even more worrying is that these under-reported results seem to be a biased sample of all results[20]. However, datasets, and effect sizes calculated using published datasets, are less likely to suffer from this kind of issue. Datasets that support published studies can be also used to verify or supplement the reported results of the study, increasing the number of effect sizes that can be calculated (missing or contradictory reported results).

Finally, meta-analyses conducted using the values reported in studies have to combine effect sizes calculated in a different way (as primary studies analyse their data and report the results differently). Effect sizes can be calculated in a consistent manner if the original data are used (such as in our case study; Box 1), thus leading to directly compatible effect sizes[18].

Limitations of open data meta-analysis

Despite the apparent benefits, our meta-analysis conducted using non-target research data suffers several limitations. These should not discourage data-driven meta-analysis, but rather be acknowledged and, if possible, adequately resolved.

First, as our case study demonstrates, the description of datasets is often insufficient to enable a sensitive and targeted search. This means that data searches may retrieve a substantial number of irrelevant datasets, while also missing some relevant ones. However, this has always been a limitation of meta-analyses, and we believe this will only improve as the scientific community continues to embrace the advised data standards (for example, ref. [21]), supported by improvements in the data curation by research institutions and scientific repositories. The second and related issue is the quality of the retrieved datasets, a number of which might need to be excluded due to the lack of sufficient information. In our case, this reduced the number of species for the analysis and led to loss of a number of taxonomic groups (panel b of the Box 1 figure). The third issue is the misinterpretation of data used in meta-analysis[16], especially when using non-target studies that addressed different questions from the proposed study. Contacting data owners is probably the best approach to address this issue (for example, we excluded 4 out of 18 datasets based on owner comments) and should thus be standard in open-data meta-analysis. The outlined issues might make meta-analysis based on data more time-consuming compared with traditional meta-analysis, but based on our experience this will vary from case to case.

Conclusion

The meta-analysis approach has become increasingly important across ecological and evolutionary research fields, having a strong impact on future research, interventions and policies. Here, we introduce a new standard on how to conduct a data-driven meta-analysis that, in contrast to the conventional meta-analysis, uses research data rather than published studies. This new standard is now possible given that the amount of open research data has been steadily increasing across evolutionary and ecological fields. We show that new questions can be addressed with the use of this ever-growing data landscape, broadening the scope of meta-analysis in evolutionary ecology. In addition, by embracing open data, evolutionary ecology has the potential to benefit from a spectrum of higher standards and reporting practices brought in the new era of open science. ❐

Antica Culina[1]*, Thomas W. Crowther[1,2], Jip J. C. Ramakers[1], Phillip Gienapp[1] and Marcel E. Visser[1]

[1]Netherlands Institute of Ecology (NIOO-KNAW), Wageningen, Netherlands. [2]Institute of Integrative Biology, ETH Zurich, Zurich, Switzerland.

**e-mail: A.Culina@nioo.knaw.nl*

Published online: 18 June 2018
https://doi.org/10.1038/s41559-018-0579-2

References

1. Cadotte, M. W., Mehrkens, L. R. & Menge, D. N. L. *Evol. Ecol.* **26**, 1153–1167 (2012).
2. Gurevitch, J., Koricheva, J., Nakagawa, S. & Stewart, G. *Nature* **555**, 175–182 (2018).
3. Jennions, M. D., Kahn, A. T., Kelly, C. D. & Kokko, H. *Evol. Ecol.* **26**, 1119–1151 (2012).
4. Stewart, G. B. & Schmid, C. H. *Res. Synth. Methods* **6**, 109–110 (2015).
5. Lortie, C. J., Stewart, G., Rothstein, H. & Lau, J. *Res. Synth. Methods* **6**, 246–264 (2015).
6. Bayliss, H. R. & Beyer, F. R. *Res. Synth. Methods* **6**, 136–148 (2015).
7. Parker, T. H. et al. *Trends Ecol. Evol.* **31**, 711–719 (2016).
8. Simmonds, M. C. et al. *Clin. Trials* **2**, 209–217 (2005).
9. Roche, D. G., Kruuk, L. E. B., Lanfear, R. & Binning, S. A. *PLoS Biol.* **13**, e1002295 (2015).
10. Wallis, J. C., Rolando, E. & Borgman, C. L. *PLoS ONE* **8**, e67332 (2013).
11. Evans, S. R. *PloS Biol.* **14**, 1–9 (2016).
12. Koricheva, J., Gurevitch, J. & Mengersen, K. (eds) *Handbook of Meta-Analysis in Ecology and Evolution* (Princeton Univ. Press, Princeton, Oxford, 2013).
13. Nakagawa, S., Noble, D. W. A., Senior, A. M. & Lagisz, M. *BMC Biol.* **15**, 18 (2017).
14. Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G. & The PRISMA Group *PLoS Med.* **151**, 264–269 (2009).
15. Culina, A. et al. *Nat. Ecol. Evol.* **2**, 420–426 (2018).
16. Mills, J. A. et al. *Trends Ecol. Evol.* **30**, 581–589 (2015).
17. Ramakers, J. J. C., Culina, A., Visser, M. E. & Gienapp, P. *Nat. Ecol. Evol.* https://doi.org/10.1038/s41559-018-0577-4 (2018).
18. Mengersen, K., Gurevitch, J. & Schmid, M. D. in *Handbook of Meta-Analysis in Ecology and Evolution* (eds Koricheva, J., Gurevitch, J. & Mengersen, K.) 300–313 (Princeton Univ. Press, Princeton, Oxford, 2013).
19. Cote, I. M. & Reynolds, J. D. *Evol. Ecol.* **26**, 1237–1252 (2012).
20. Cassey, P., Ewen, J. G., Blackburn, T. M. & Møller, A. P. *Proc. R. Soc. Lond. B* **271**, 451–454 (2004).
21. Wilkinson, M. D. et al. *Sci. Data* **3**, 160018 (2016).
22. Wood, C. & Brodie, E. *Ecol. Lett.* **19**, 1189–1200 (2016).
23. Husby, A., Visser, M. E. & Kruuk, L. E. B. *PLoS Biol.* **9**, e1000585 (2011).
24. Wilson, A. J. et al. *PLoS Biol.* **4**, e216 (2006).

Author contributions

A.C. collected the data and wrote the majority of the manuscript; T.W.C., J.J.C.R, P.G. and M.E.V. all contributed to the discussion and the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-018-0579-2.