*Special Issue—Data Sharing in Ecology and Evolution***A response to Poisot et al.: Publishing your dataset is not always virtuous****Angela T. Moles¹, John B. Dickie, and Habacuc Flores-Moreno**

Angela T. Moles (a.moles@unsw.edu.au), Evolution and Ecology Research Centre, The University of New South Wales, NSW 2052, Australia

John B. Dickie (j.dickie@kew.org), Royal Botanic Gardens Kew, Wakehurst Place, Ardingly, West Sussex, RH17 6TN, United Kingdom

Habacuc Flores-Moreno (habacucfm@yahoo.com), Evolution and Ecology Research Centre, The University of New South Wales, NSW 2052, Australia

¹Corresponding author

Poisot et al. (2013) present an interesting paper that extols the benefits of researchers making their data publicly accessible. We absolutely agree that making your own primary data publically available is a virtuous, helpful, positive thing to do, and should be encouraged. However, we believe that Poisot et al. (2013) have overlooked two important factors: 1) the ethics of publishing data that were gathered by other people, and 2) the potential for enforced data release to actually slow our progress in science by removing incentives for scientists to undertake large data compilation efforts.

Publishing other people's data is not virtuous

One way in which ecologists have been making substantial progress is in compiling very large datasets from a range of sources and locations, and asking what big patterns emerge (Díaz et al. 2004, Wright et al. 2004, Koele et al. 2012). There are important reasons why these multi-owner datasets can't be published or released to the general public. First, when a data owner is asked for access to their data for a particular project, it is common for them to give permission on the conditions that (1) the data user does not pass the data on to others or publish the data, and (2) the data user only uses the data for the proposed project (in other words, the data user must ask the data owner for permission again if they wish to use the data for a different project).

Data sharing agreements along these lines are standard at all levels from individual data owners, to institutions and government agencies (e.g. the New Zealand Ministry for Environment's LUCAS database), and for data-sharing initiatives such as TRY (Kattge et al. 2011). Thus, publishing or releasing the dataset from most large data syntheses would be in direct violation of data-sharing agreements with data owners.

There is also a moral reason not to publish other people's data. If a data user takes a dataset, adds data, and publishes the resulting larger dataset, the original data contributor will get a single citation, and all future citations and use of the database would be through the more recent paper and database. That is, the people who did the original work would be deprived of all future credit. There are obviously shades of grey here. One could argue that publishing a compilation of single datapoints from studies that were performed to test specific hypotheses in specific systems allows these data to be used in novel ways, thus adding value to the original papers. However, taking major databases that have taken years of work to compile (e.g. the Kew Seed Information Database) and adding a small amount of new information is clearly a different issue.

Requirements for authors to publish their raw data currently prevent the publication of syntheses of data belonging to multiple researchers in several high profile ecology journals, unless the lead author talks all of their

data contributors in to allowing their data to be published with someone else as lead author. It is currently unclear why data owners would do this.

The view from the data-owner's side of the fence

It is common for people analysing data that they did not collect to overlook the huge amount of work that has gone in to collecting and/or collating the data. Most ecologists have enough first-hand experience that they understand the enormous amount of effort that goes in to collecting field data or conducting experiments in controlled environments. However, it is common for people to assume that collating data is a trivial task that is not worthy of recognition. This is simply not the case. The TRY database has so far taken more than 17,000 person-hours to compile (pers. comm. Jens Kattge). Similarly, the Royal Botanic Gardens Kew Seed Information Database is the result of at least 5500 person hours over the last 12 years. These estimates do not even include the time required for the primary authors to collect and publish their data. Asking the people who have made such wonderful resources available for ecologists to waive their rights to credit for their efforts would surely discourage researchers from embarking on such projects, and a lack of any concrete outputs would surely discourage funding agencies from supporting efforts to synthesise datasets. That is, a well-intended scheme along the lines of that proposed by Poisot et al. (2013) could actually substantially hinder the progress of comparative ecology.

Another problem that is often underestimated is the difficulty of standardising ecological data. Gene sequence data can be standardised relatively easily. However, ecological traits can be measured in all sorts of different ways, in different units, and using different definitions, on plants/organs in all sorts of different environments, at different times of year, and at different stages of development, and the results will often depend on the duration of the study. Further, there are serious problems associated with standardising the nomenclature of data gathered in different countries in different years. Although people are working on this problem (e.g. the Knowledge Network for Biocomplexity, <http://knb.ecoinformatics.org/index.jsp>), synthesising data from different sources is not a trivial matter.

Poisot et al. (2013) establish a good case for the benefits to science that would come from researchers making their data available. However, they state that “the contribution of data will become less and less of a criteria for authorship...thus allowing to name as authors only those who analysed the data.” They also suggest that the solution to data availability is “each research group can keep and take care of its own local database.” That is, Poisot et al. (2013) are proposing that the people who have done the hard work of

painstakingly collecting and collating datasets should do a lot more work (including paying to make the data available on a website, curating the database, and fielding requests and questions from data users), for minimal credit. Providing such information is undoubtedly good for our science, and will generate a certain amount of goodwill among the ecological community. However, this may not be sufficient motivation in the current scientific climate. Unsurprisingly, there is currently a vast gap between the proportion of scientists who would like to use other people's data, and the proportion of scientists who have made their own data available (Tenopir et al. 2011).

Encouraging data sharing is going to require something more proactive than finger-wagging data owners about moral obligation. It will require us to provide concrete incentives for data contribution, and to make it as easy as possible for data owners to make their data available. The National Science Foundation of the United States' switch to listing “products” rather than “publications” on grant applications (Piwowar 2013) is a small step in the right direction. However, a high proportion of the citations of data sources are in the appendices of papers, where they are not visible to citation-tracking software. Further, each dataset will be obsolete as soon as someone publishes a version with extra data. Thus, published datasets seem unlikely to be high points of many people's research output.

Co-authorship is a useful and widely-recognised currency in science, with benefits to both data owners and data users. It is not clear to us what the downside is of giving authorship to the people who did the hard work of collecting the data. After all, interpreting analyses and writing manuscripts without the involvement of the people who actually know the details of data collection and understand the natural history of the study systems from which the data were gathered is a precarious business.

References

- Díaz, S., J. G. Hodgson, K. Thompson, M. Cabido, J. H. C. Cornelissen, A. Jalili, et al. 2004. The plant traits that drive ecosystems: Evidence from three continents. *Journal of Vegetation Science* 15:295–295.
- Kattge, J. and S. Díaz and S. Lavorel and I. C. Prentice and P. Leadley and G. Bönsch, et al. 2011. TRY - a global database of plant traits. *Global Change Biology* 17:2905–2935. [CrossRef](#)
- Koele, N., I. A. Dickie, J. Oleksyn, S. J. Richardson, and P. B. Reich. 2012. No globally consistent effect of ectomycorrhizal status on foliar traits. *New Phytologist* 196:845–852. [CrossRef](#)
- Piwowar, H. 2013. Altmetrics: Value all research products. *Nature* 493:159–159.

- Poisot, T., R. Mounce, and D. Gravel. 2013. Moving towards a sustainable ecological science: don't let data go to waste! *Ideas in Ecology and Evolution* 6: 11–19 [CrossRef](#)
- Tenopir, C., S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, et al. 2011. Data sharing by scientists: practices and perceptions. *PLoS ONE* 6:e21101-e21101. [CrossRef](#)
- Wright, I. J., P. B. Reich, M. Westoby, D. D. Ackerly, Z. Baruch, F. Bongers, et al. 2004. The worldwide leaf economics spectrum. *Nature* 428:821-827. [CrossRef](#)

Subject Editor: Karthik Ram