# Recognising the value of the scientific resources generated by data collectors and code developers

Robert. M. Ewers[1,*], Jos Barlow[2], Cristina Banks-Leite[1] and Carsten Rahbek[1,3]

[1] Imperial College London, Silwood Park Campus, Buckhurst Road, Ascot SL5 7PY, United Kingdom

[2] Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, United Kingdom

[3] Center for Macroecology, Evolution and Climate, Natural History Museum of Denmark, University of Copenhagen, 2100 Copenhagen, Denmark.

* Email: r.ewers@imperial.ac.uk

The current, authorship-based system for recognising individual contributions to science only patchily recognises the contributions of the primary data collection that underpins, and code development that supports, the entire discipline. While data collectors and code developers – scientific resource generators – are progressively being forced to donate the grant income, time and effort of generating, curating and documenting data and code to the discipline as a whole [1-3]. Resource users – those that re-use previously published data and codes to generate new knowledge and publications – benefit from that time and effort but are not required to recognise it in any standardised manner. We need a new way to quantify and value what is currently anonymous; the fundamental contribution to scientific progress that generating scientific resources provides.

Many scientists agree that authorship is the ultimate reward for collecting data or developing code. However, the Vancouver Protocol tellingly states that "*Participation solely in the … collection of data does not justify authorship.*" Citations are routinely raised as the obvious approach to solving this dilemma [4, 5], but it is not enough. Citations carry less value to a scientist than authorship. Moreover, citations to scientific resources are agnostic to the impact of the papers that used those resources, resource citations are commonly buried in supplementary material where they do not get picked up by citation tracking software, and published resources not associated with a published manuscript do not contribute to a scientists' citation indices.

We suggest one solution is to divorce authorship of a manuscript from authorship of the resources used in the manuscript, which can be achieved by creating separate categories of authorship: manuscript and resource authors. Here, a published paper would come with two separate author lists. Manuscript authors are those who developed the question, analysed and interpreted the data, and wrote the paper; "*authorship for authors*" [6]. Resource authors are those who contributed some or all of the data that was analysed or code that was used. Membership of the two author lists need not be mutually exclusive, as a single person could reasonably contribute resources and contribute to the manuscript. In this system, a resource generator can still receive credit for contributing to the paper, without implying they agree with, understand, or have even seen, the analysis and the conclusions the manuscript authors have presented.

Resource authorship provides a path to quantify the value of a scientist's provision of resources to the wider community, and could be implemented within the framework of the existing, citation-based recognition system. Resource contributions could reasonably be tracked through the use of exactly the same citation indices already in widespread use, but applied to resource rather than manuscript authorship. This would ensures scientists contributing data or code that are frequently re-used in highly cited, influential papers will have higher resource citation metrics than those contributing resources that are infrequently used and published in low impact papers.

Separating the impact of generating scientific resources from the impact of using those resources provides a way out of the resource generator-resource user tension. The two are complementary aspects of a shared scientific enterprise. Data and reproducible codes represent empirical truth; quantitative, repeatable measurements of the world around us against which we test our understanding. The papers we write are our qualitative interpretation of what those data and codes tell us; they are ephemeral position statements that implicitly embed the sum of our experiences, knowledge and biases to date. Both are important contributions to the advancement of science, and both need to be represented when quantifying the contribution that individuals make to that advance.

**References**


1. Whitlock, M.C. (2011) Data archiving in ecology and evolution: best practices. Trends in Ecology & Evolution 26 (2), 61-65.

2. Mislan, K.A.S. et al. (2016) Elevating the status of code in ecology. Trends in Ecology & Evolution 31 (1), 4-7.

3. Peng, R.D. (2011) Reproducible research in computational science. Science 334 (6060), 1226-1227.

4. Amann, R.I. et al. (2019) Toward unrestricted use of public genomic data. Science 363 (6425), 350-352.

5. Pierce, H.H. et al. (2019) Credit data generators for data reuse. Nature 570, 30-32.

6. Baskin, T.I. (2018) Keep authorship for writers. Nature 562, 494.