Archiving primary data: solutions for long-term studies

Mills, James A; Teplitsky, Céline; Arroyo, Beatriz; Charmantier, Anne; et al; Griesser, Michael; et al; Zedrosser, Andreas

Abstract: The recent trend for journals to require open access to primary data included in publications has been embraced by many biologists, but has caused apprehension amongst researchers engaged in long-term ecological and evolutionary studies. A worldwide survey of 73 principal investigators (PIs) with long-term studies revealed positive attitudes towards sharing data with the agreement or involvement of the PI, and 93% of PIs have historically shared data. Only 8% were in favor of uncontrolled, open access to primary data while 63% expressed serious concern. We present here their viewpoint on an issue that can have non-trivial scientific consequences. We discuss potential costs of public data archiving and provide possible solutions to meet the needs of journals and researchers.

DOI: https://doi.org/10.1016/j.tree.2015.07.006

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-120573
Accepted Version

Originally published at:
DOI: https://doi.org/10.1016/j.tree.2015.07.006
Solutions for archiving data in long-term studies - a reply to Whitlock et al.


*JAM and CT contributed equally

Authors’ affiliations
1 10527A Skyline Drive, Corning, NY 14830, USA
2 CESCO, UMR7204 Sorbonne Universités-MNHN-CNRS-UPMC, CP51, 55 Rue Buffon, 75005 Paris, France
3 Centre d’Ecologie Fonctionnelle et Evolutive UMR 5175, Campus CNRS, 1919 Route de Mende, 34293 Montpellier, Cedex 5, France
4 Instituto de Investigacion en Recursos Cinegeticos (IREC) (CSIC-UCLM-JCCM), Ronda de Toledo s/n, 13005 Ciudad, Real, Spain
5 Institute of Avian Research, “Vogelwarte Helgoland”, An der Vogelwarte 21 D26386 Wilhelmshaven, Germany
6 Department of Animal and Plant Sciences, The University of Sheffield, UK
7 Institute of Biological and Environmental Sciences, University of Aberdeen, UK
8 Department of Ecology and Evolutionary Biology, University of California, 621 Young Drive South, Los Angeles, CA 90095-1606, USA
9 CNRS,Université Lyon 1, Université de Lyon, UMR 5558, Laboratoire Biométrie et Biologie Évolutive, 43 boulevard du 11 Novembre 1918, F-69622 Villeurbanne, Cedex, France
10 Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2E9
11 Department of Vertebrate Zoology, Faculty of Biology, Lomonosov Moscow State University, Leninskii Gory 1/12, 119234 Moscow, Russia
12 UMR 5174 EDB Laboratoire Évolution et Diversité Biologique, CNRS, ENFA, Université Toulouse 3 Paul Sabatier, 31062 Toulouse Cedex 9, France
13 Department of Evolution, Ecology & Genetics, Research School of Biology, The Australian National University, Canberra, ACT, Australia
14 Département de biologie & Centre d’études nordiques, Université Laval, 1045 avenue de la Médecine, Quebec (QC), G1V 0A6, Canada
15 29 St Mary’s Close, Shincliffe, Durham, DH1 2ND, UK
16 Centre for Ecology & Hydrology, Bush Estate, Penicuik, EH26 0QB UK
17 Behavioural Ecology, Department of Biology, Ludwig-Maximilians University of Munich, Planegg-Martinsried, Germany
18 Evolutionary Ecology of Variation Research Group, Max Planck Institute for Ornithology, Seewiesen, Germany
19 Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, A.P. 70-275, México D.F. 04510, México
20 Technical Resource Branch, Saskatchewan Ministry of Environment, 3211 Albert Street, Regina, Saskatchewan, Canada S4S 5W6
In our recent paper [1], we discussed some potential undesirable consequences of public data archiving (PDA) with specific reference to long term studies and proposed solutions to manage these issues. We reaffirm our commitment to data sharing and collaboration, both of which have been common and fruitful practices supported for many decades by researchers involved in long-term studies. We acknowledge the potential benefits of PDA [e.g. 2], but believe that several potential negative consequences for science have been underestimated [1, see also 3,4]. The objective of our recent paper [1] was to define practices to simultaneously maximize the benefits and minimize potential unwanted consequences of PDA.

Commenting on our paper, several former and current editors of major ecology and evolution journals [5] acknowledge the need to improve data archiving practices to account for the concerns presented in [1]. The fact that editors of several journals were willing to comment on our paper underlines the importance of this issue, and we are keen to continue this dialogue to identify potential solutions. Following our [1] and Roche et al.’s [6] suggestions, Whitlock et al. [5] endorse as good practice longer embargos (5 years) and encourage cooperation or collaboration with data providers. Both steps are major advances as many of the Principal Investigators (PIs) in [1] have been denied longer embargos, and the practice of consulting PIs to ensure that data files are properly interpreted is not a formal policy in any scientific journal.

We welcome these positive developments but underline three concerns, two of which extend beyond the purview of individual journals:

Whitlock et al. [5] mention that current policies “require only that authors make available the data necessary to recreate the analyses and results in the published manuscript”. For an article that includes an analysis based on a pedigree and individual data, or on lifetime reproductive success and potential predictor variables, this requirement involves providing a detailed data base of the breeding performance of individuals and their progeny over decades. The costs of data gathering, including resources beyond monetary ones, are borne by the data providers and their institutions not by those who would use the data; consequently providing such extensive datasets is sustainable if the data are used only to verify the original analysis. Extending an embargo to five years for such data is a good step, but for studies that extend over decades, a longer embargo is warranted, notably to further encourage potential users to contacts PIs to get the latest version of the data, and ideally collaborate.

Databases from long-term studies are an evolving infrastructure that underpin numerous publications. New data are added each year, and errors and omissions are corrected regularly. Over time, archives often include various versions of fragmented datasets which [(i)] could be combined by
others in ways that the data collectors were already doing or planning to do themselves, or (ii) may differ from each other in ways that are likely to lead to misinterpretation of the data. A single journal’s PDA policy cannot ensure that data from long-term studies are not misused. It must be a community decision. Some potential solutions include archiving at institutional servers with separate policies for the distribution of data necessary to reproduce previously published analyses and data requests for additional analyses. The additional analyses would require collaboration with the PI.

Finally, journal editors do not control the policies of funding agencies, but their stature in the community can be influential. Whitlock et al. [5] suggest that funders should set standards for openness. However, long-term studies typically involve several grants and multiple funding agencies, sometimes from different countries. Hence, any discrepancy between their policies can lead to potentially insoluble conflict. Institutions that fund a significant proportion of the research, potentially over decades, may also question the value of continued funding if the data are freely available to individuals from other organizations.

We are encouraged by the letter from Whitlock et al. [5], but believe that there are additional issues that need to be addressed. Some of these may be solved by a more explicit and flexible policy on longer embargos, data storage on institutional servers and involvement of the principal investigators in new analyses using the data they produced, through collaboration or reviews. We hope that this important dialogue will continue.

Cited references


6. Roche, D.G. et al. (2014) Troubleshooting public data archiving: Suggestions to increase
participation. *PLoS Biol.* 12, e1001779