Cell PRESS

Data archiving in ecology and evolution: best practices

Michael C. Whitlock

Department of Zoology, University of British Columbia, Vancouver, BC, Canada, V6T 1Z4

Many ecology and evolution journals have recently adopted policies requiring that data from their papers be publicly archived. I present suggestions on how data generators, data re-users, and journals can maximize the fairness and scientific value of data archiving. Data should be archived with enough clarity and supporting information that they can be accurately interpreted by others. Re-users should respect their intellectual debt to the originators of data through citation both of the paper and of the data package. In addition, journals should consider requiring that all data for published papers be archived, just as DNA sequences must be deposited in GenBank. Data are another valuable part of the legacy of a scientific career and archiving them can lead to new scientific insights. Archiving also increases opportunities for credit to be given to the scientists who originally collected the data.

Data archiving

The foundation of science is data; that is, the collection of information about the natural world obtained through experiment and observation. Without data, there can be no science. Yet most of the data collected, particularly in ecology and evolutionary biology, is quickly lost to science. Other than through summaries posted in subsequent publications, most data are never accessible to anyone other than their original collectors, and many data are eventually lost, even to their collector, over the course of time. Recently, scientists have collectively become more aware of the value of data and of the importance of data preservation [1,2]. In particular, ecology and evolutionary biology journals [3–8] and funding agencies (Box 1) have recently adopted policies that either encourage or require data archiving as part of the publication process, such as the Joint Data Archiving Policy being adopted by several journals [3]. Powerful new data repositories, such as Dryad (http://datadryad.org) and KNB (http://knb.ecoinformatics.org), have enabled all kinds of data in ecology and evolution to be archived. These are in addition to repositories for more specialized data, such as DNA sequences (GenBank; http://www.ncbi.nlm.nih.gov/genbank/), phylogenetic trees (TreeBASE; http://www.treebase.org/), microarrays (GEO; http://www.ncbi.nlm.nih.gov/geo/), vegetation plots (VegBank; http://www.vegbank.org/), and hydrology data (CUAHSI-HIS; http://his.cuahsi.org/), among others, including projects that emphasize the interoperability of archives (e.g. DataONE; https://www.dataone.org/).

Data archives serve science in a variety of ways (Box 2). Publicly archived data enable more transparent science, with better error checking and verification of results. Archiving also enables data to be re-used for broader meta-analyses and to address new questions. Available data can serve a powerful educational role, both in teaching the statistical and technical aspects of research and to engage students in the process of science. Public data archiving is also a powerful mechanism for data security, providing a mechanism by which data can be saved and re-accessed by the original authors and others even after hard disk failure or other catastrophes.

The broad spectrum of possible use and re-use of archived data speaks to an array of potential users of the archive, including the original collectors of the data, their collaborators, interested scientists, reviewers, meta-analysts, students, government agencies, funding councils, or the general public, who fund most of the research. For such a spectrum of users to get the most value from an archived data set, it is important that data be archived in a useable format that has a high probability of being interpretable by others both now and in the future. Some simple guidelines can increase the usability of data. Because archiving is relatively new to many ecologists and evolutionary biologists, I briefly consider features that make an archived data set most valuable.

For data to be archived, it also needs to be clear that the original collectors of the data will get appropriate credit for their work [2]. As a field, ecologists and evolutionary biologists are at an early stage in a cultural shift about data re-use. It is necessary to start thinking about the various ways in which researchers, both collectively and individually, should credit the originators of valuable data.

Finally, as a new era begins in the relation of the ecology and evolutionary biology fields to data, one needs to consider what journals and editors might do to make this transition to data archiving most valuable and in the best interests of science. Journals can have a powerful role in facilitating data archiving and in giving credit for uses of data.

In this review, I describe some of the issues, both technical and cultural, that need to be addressed with the move to archiving more data. I make some practical suggestions, based more on pragmatism than idealism, for the ways that science might be best served by the actions of original authors, data re-users, and journals. I label these suggestions as 'best practices' to emphasize that they are

Corresponding author: Whitlock, M.C. (whitlock@zoology.ubc.ca).

  doi:10.1016/j.tree.2010.11.006

Box 1. Examples of funding agencies with recently adopted data archiving policies

- National Science Foundation (2004): NSF Grant Proposal Guide (GPG) (NSF 04–23) (http://www.nsf.gov/pubs/gpg/nsf04_23/6.jsp).
- National Institutes for Health (2003): Final NIH Statement on Sharing Research Data (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html).
- Biotechnology and Biological Sciences Research Council (2010): BBSRC Data Sharing Policy (http://www.bbsrc.ac.uk/web/FILES/Policies/data-sharing-policy.pdf).
- Natural Environment Research Council (2010): NERC Data Policy (http://www.nerc.ac.uk/research/sites/data/policy.asp).

not all hard-and-fast rules. I emphasize that, by using the word 'best', I do not mean to imply perfection: the cultural, technical and publishing practices that are used will evolve over time as more experience is gained in these areas. These thoughts are meant as one starting point on that evolution.

I focus on the general kinds of data that are associated with many ecology and evolutionary biology papers of the type that can be placed in an archive such as Dryad or KNB. Being broad, journals in these fields and related data archives need to cope with data formats that are *sui generis* and not standardized. In the context of the new data policies of ecology and evolutionary biology journals, I also focus on the issues that arise with archiving data associated with papers (rather than, say, data that are associated with unpublished projects). Data sets associated with publications have the advantage that the associated paper already gives much of the important context, methods and meta-data required to interpret the results accurately.

Best practices for data creators

The most important step in data archiving for data originators is to decide to archive data, in a public and long-lasting forum. Without public archiving, most data are quickly lost to science [9]. If data are archived in a conscientious fashion, their value to science and to their originators can increase dramatically over time.

Choose an archive that is most suitable for your type of data. For example, GenBank is of course the right place for DNA sequence data; TreeBASE is the right place for phylogenetic trees and the data matrices used to generate them; and archives such as GEO support microarray, next-generation sequencing and other forms of high-throughput functional genomic data. Other data have multiple possible hosts. All data in the fields of ecology and evolutionary biology can be archived at the Dryad repository or KNB, provided there is not an established site for that kind of data. In any case, put the data in a place that will stand the test of time; data placed on personal websites tend not to be available for long [10].

The central goal to have in mind when archiving your own data is to ensure that a new user, perhaps someone unknown to you working with the data 20 years later, can correctly interpret the results and derive correct conclusions from the data. Just as it is important in scientific communication to be clear and precise when writing papers, it is also crucial to communicate effectively about data and their context. Such clarity not only improves the chance that the data will be used to generate correct new insights into the natural world (and that someone cites you), but it also reduces the number of avoidable questions that you might have to answer from other scientists about the data.

Box 2. Why archive data?

Data are, in many cases, a precious scientific resource, and their value is enhanced, not exhausted, by the first publication of conclusions drawn from them. When data are publicly archived, science and scientists benefit in several ways.

Verification of published results

Unfortunately, many published papers report on data that have been analysed incorrectly, and 5–10% of papers might have major conclusions that are not supported by their data [24–26]. If the data are available for analysis by others, important mistakes can be caught and corrected. Intentional scientific misconduct is probably less common [27], but public data archiving provides a powerful method for preventing and correcting misconduct [28]. There will be a powerful incentive for data to be checked more carefully and analyzed more appropriately before publication when scientists know that their data will be available for review.

Better meta-analysis

Meta-analysis has rapidly developed as a tool in ecology and evolutionary biology, increasing 40-fold over the past two decades (based on a Web of Science search, June 2010; http://apps.isiknowledge.com/). However, meta-analysis requires accurate representations of results and, often, the necessary information is not provided in published papers. Archiving provides meta-analysts increased access to the information needed for maximal accuracy and precision. Papers are more likely to be included in future meta-analysis if full quantitative information is available.

New questions

Archived data will in some cases result in answers to novel questions not considered by the original authors. Bumpus' [29] classic data set on house sparrow survival from 1898 has been used countless times, often for reasons never considered by the original author, including assaying multivariate selection. All data from the Hubble space telescope are archived and, through this archive, new extrasolar planets have been discovered without the expense of new observation [30].

Increased citation and credit

When data are archived, their authors are given more credit by the scientific community in ways that can translate to greater career success. Papers that archive data publicly are cited 69% more often than papers that withhold the data [31]. Moreover, archived data can be cited in the same way as papers, and citation of both the paper and the data package should be encouraged. Authors that archive valuable data will have papers that are cited more often, and they will also accrue citations to their data sets. With data archiving, the legacy of a scientific career will be measured not only by an author's papers, but also by the value of the data that they contribute.

New opportunities for teaching and learning

Data can be extremely valuable for case studies, as teaching tools for learning the process of science and biostatistics.

Reducing loss

Public data archives also reduce the risk that the originators of the data will lose access to their own data through hard-disk failure or through losing information about the context of their collection. If data is publicly archived, their original author(s) will have less to fear about data loss.

If the data are associated with a publication, the paper will probably already convey many of the methodological details needed to use the data again. However, many important details might be left out of the paper, including the meaning of column headers in the data file, units, precise localities, indicators for missing data, codes for categorical variables, and so on. A short, clear readme file attached to each package of archived data should clear up such remaining details. Even better, metadata can be recorded using a standard format, such as EML (ecological metadata language; http://knb.ecoinformatics.org/software/eml/) for greater re-usability. Archived data have less value if the metadata required to interpret them are not clear (University of Texas Libraries (2010) Recommended file formats. http://repositories.lib.utexas.edu/recommended_file_formats).

Start data management while analyzing the data and writing the paper. It is much easier to describe the data set correctly while it is fresh in your mind. The data should be archived at a level ready for a statistical analysis program, and should be given at the individual level. For example, a behavior trial in a Y-maze might have been videotaped; the archived data should record the choice made by the fish and (if relevant) the time taken, but not necessarily the movie file itself.

Some kinds of data ought not to be archived or, at the very least, significant care must be taken to anonymize some sensitive details. Information such as the location of populations of endangered species or culturally significant archaeological sites should, in many cases, be left out of archived information. Data from human subjects are especially sensitive, and care must be taken to maintain the appropriate level of patient privacy (for guidelines on this issue, see [11] and National Human Subjects Protection Advisory Committee (2002) http://www.aera.net/humansubjects/NHRPAC_Final_PUDF.pdf).

Whenever possible, use a non-proprietary file format that is more likely to be readable in the future. For example, a text file is better than a Word .doc file, and comma-delimited text is better than an Excel file. No matter how long lasting a particular computer program might seem, programs come and go, and a data set stored in an obsolete format will be difficult to read. Non-proprietary formats are also readable by a larger diversity of programs. Suggestions for better file formats are available (University of Texas Libraries (2010) Recommended file formats. http://repositories.lib.utexas.edu/recommended_file_formats), and other good suggestions about data management techniques that make files easier to use are given by Borer *et al*. [12].

After creating the data and readme files, try to rerun several of the analyses reported in the paper. Others might try to replicate the analysis, and it will save everyone concerned a lot of time if the file unambiguously includes the right data. Just as with a paper, it can be useful to have a friend or colleague look over the readme file with the data.

Deciding when to archive data can sometimes be straightforward, and sometimes challenging. Data from a complete, stand-alone study can be archived immediately upon publication of the results, or even sooner, without cost to their originators. Other studies might be ongoing, and not all aspects of the study might have been reported in the first paper. In such cases, the original authors might worry about others' access to the data before they have fully published the work. Several provisions in the data archiving policies of most journals have been implemented to balance between such concerns and scientific openness. Most journals allow, and Dryad implements, the option of a one-year embargo after publication on public access to the data. This results in the option of more than a year for subsequent analysis by the original authors after submission of the first paper and before public scrutiny of the data. Moreover, the journal archiving policies only require the data that are necessary to recreate the results of the published paper, not the full data set from the project. Other data collected in the same project need not be archived until it is used. Finally, most journal policies allow for strong editorial discretion; special cases might well merit longer embargo periods. Many advantages accrue to science from immediate archiving, and shorter lags to access are preferable. However, results in the fields of ecology and evolutionary biology have long-term value; data must be preserved for posterity even if they are not immediately accessible.

In any case, it is preferable to archive the data soon after their collection and analysis, while the details of their interpretation and the metadata are still available. Even if the data are embargoed from public access until a later date, the important information can be saved while it is still readily available.

Best practices for data users

When researchers reuse data that have been archived by previous workers, it is crucial that they respect the work required to create those data, that they remember the greater insight into the context of those data that their originators will have, and that they properly acknowledge the debt to those original authors. Part of respecting the original authors is to gain as much understanding of the methods of the original work as possible; therefore, data should never be reused without careful reading of the original papers and associated materials. Read as broadly as possible about the study system to minimize the possibility of misuse. To ensure that the data are archived correctly and that you understand that data, it is always good practice to first try to recreate some results from the original paper before proceeding to use the data for new purposes.

The researchers who collect the data know the methods and context of the data collection better than can be communicated in a paper or meta-data; therefore, it is always wise to contact the original authors to discuss the use of the data. In many cases, those authors will be able to help substantively with the new interpretation of the data and, when that new input reaches a non-trivial level, the original author(s) should be offered co-authorship on the new project. Co-authorship should not, however, be necessary by default, as the originators of the data will gain credit by citation in the new work. Out of respect to the crucial work done by the original collectors of the data, err on the side of generosity with authorship.

At the very least, whenever data are reused, researchers must cite not only the paper or papers in which they were originally described, but also the data package itself, using the norms suggested by the data archive or journal. It is

especially useful if the citation is formatted in a similar way to a paper citation, as part of the 'Literature cited' section. For example, here is a citation to recent work from my lab [13,14]. The data archived in Dryad, for example, has been released under a Creative Commons Zero (CC0) license, which releases the data legally into the public domain. This does not free re-users from complying with established scientific cultural norms about giving credit through citation, just as is already the case with citation of previous scholarly work presented in papers. Cite others' data as you would like your data to be cited.

If there seems to have been an error in the previous analysis of the data, check and double check your results, and then contact the original authors for clarification. If the error is substantial enough to change a major conclusion, it might be appropriate to contact the journal with a corrigendum. If this can be done with the cooperation of the original authors, the process will be more straightforward.

Best practices for editors and publishers

Journals have a vital role in data archiving in at least two ways. They can encourage or require data archiving, and they can help ensure that archived data is used in a way that is consistent with the overall aims of science.

Many journals in ecology and evolution have adopted policies that either encourage or require data archiving as part of the publication process. Journals already have a gatekeeper role for facilitating publication of good science; they can and should also have a leadership role in improving science by creating incentives for data archiving. Several journals in evolutionary biology have already banded together to create a shared policy to require data archiving as a condition for publication [3–7]; other journals are invited to join this initiative.

Many other journals have policies that require data sharing upon request. These policies are a great first step, but they are insufficient to achieve reliable preservation of data. If data are not publicly and collectively archived, they are likely to be lost over time, as the original authors change careers or employers, or as data are lost owing to hard-drive failure or death of the scientists [9]. A central public archive preserves data over the long term in a coordinated manner; moreover, if the data are archived while still fresh in the minds of their creators, they are more likely to be placed in a context within which they can be reused and understood. Unfortunately, even for papers published in journals with data policies that require authors to share upon request, the rates of data sharing are low [10,15–20]. Moreover, if data are deposited before publication, compliance with sharing policies is more easily assured. Many journal editors report draining experiences dealing with authors who submitted papers under clear data-sharing policies but who subsequently refused to share those data when contacted by other scientists. If data are publicly archived, subsequent access is assured, without extra involvement from journals or funding agencies.

Some journals choose to archive data in their own online supplemental materials. This is better than no public access to data, but it is less desirable than central archives for several reasons. Supplemental material links decay over time [20–22], and supplemental material is typically not curated to the same standard as provided by archives. Indexed archives offer greater discoverability of the data sets, and they make it easier for later users to find the data. Because the data sets can be linked to the original papers, these search functions also provide a valuable path for new users to find the papers of a particular journal.

Journals also have a key role in establishing cultural norms for data reuse. As a first step, journals should facilitate the citation of data, using standard bibliographic formats when available. For example, Dryad data sets are all assigned a DOI (digital object identifier), which enables citations to data to be tracked in the same way as citations to papers. When an author reuses data from a previous publication, both the original paper and the data set should be cited. In this way, the authors are given credit not only for their intellectual conclusions presented by the paper, but also for the contribution that the data set itself represents.

When a paper is submitted that makes extensive use of a particular data set, it seems wise to have a policy that the original authors of that data set be invited to provide a review. In this way, any misuse of the data has a high probability of being caught and corrected, and the original authors are also made aware (if they have not been before) about the use of the data that they collected.

Conclusion

As the fields of ecology and evolutionary biology move towards a more open approach to science, encouraging or even expecting archiving of data, the scientific culture must shift in some small and larger ways. Researchers already value data as the foundation of their sciences, but they must also reward the collection of useful data as an achievement in its own right [23]. By carefully documenting and preserving their own data in a permanent archive, researchers can gain a greater and more cost-effective understanding of nature. By creating a culture of giving credit by full citation to the originators of data in subsequent work, researchers can more appropriately acknowledge the contribution of the data itself and, in so doing, create a valuable path towards increasing the legacy of each scientist's work. Data collectors, data users, journals, editors, publishers, and even deans and promotion committees all have important roles in facilitating the transition of the ecology and evolutionary biology fields into one that gives more respect to the legacy of researchers' data.

Acknowledgments

This article has benefited greatly from conversations with, and comments by, Allen Moore, Heather Piwowar, Todd Vision, Mohamed Noor, Daphne Fairbairn and Sally Otto, as well as the entire Management Board of Dryad and anonymous reviewers. MCW is funded by the Natural Science and Engineering Research Council (Canada).

References

1 Vickers, A.J. (2006) Whose data set is it anyway? Sharing raw data from randomized trials. *Trials* 7, 15
2 Costello, M.J. (2009) Motivating online publication of data. *BioScience* 59, 418–427
3 Whitlock, M.C. *et al.* (2010) Data archiving. *Am. Nat.* 175, 145–146
4 Rausher, M.D. *et al.* (2010) Data archiving. *Evolution* 64, 603–604

5 Moore, A.J. *et al.* (2010) The need for archiving data in evolutionary biology. *J. Evol. Biol.* 23, 659–660
6 Rieseberg, L. *et al.* (2010) Editorial and retrospective 2010. *Mol. Ecol.* 19, 1–22
7 Butlin, R. (2010) Data archiving. *Heredity*, DOI: 10.1038/hdy.2010.43
8 Uyenoyama, M.K. (2010) MBE editor's report. *Mol. Biol. Evol.* 27, 742–743
9 Michener, W.K. *et al.* (1997) Nongeospatial metadata for the ecological sciences. *Ecol. Appl.* 7, 330–342
10 Wren, J.D. (2008) URL decay in MEDLINE – a 4-year follow-up study. *Bioinformatics* 24, 1381–1385
11 Hrynaszkiewicz, I. *et al.* (2010) Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. *Trials* 11, 9
12 Borer, E.T. *et al.* (2009) Some simple guidelines for effective data management. *Bull. Ecol. Soc. Am.* 90, 205–214
13 Yeaman, S. *et al.* (2010) Data from: no effect of environmental heterogeneity on the maintenance of genetic variation in wing shape in *Drosophila melanogaster*. *Dryad Dig. Repository* DOI: 10.5061/dryad.1719
14 Yeaman, S. *et al.* (2010) No effect of environmental heterogeneity on the maintenance of genetic variation in wing shape in *Drosophila melanogaster*. *Evolution* 64, 3398–3408
15 Noor, M.A.F. *et al.* (2006) Data sharing: how much doesn't get submitted to GenBank? *PLoS Biol.* 4, e228
16 Savage, C.J. and Vickers, A.J. (2009) Empirical study of data sharing by authors publishing in PLoS journals. *PLoS ONE* 4, e7078
17 Campbell, E.G. *et al.* (2002) Data withholding in academic genetics: evidence from a national survey. *JAMA* 287, 473–480
18 O'Leary, F. (2003) Is email a reliable means of contacting authors of previously published papers? A study of the *Emergency Medicine Journal* for 2001. *Emerg. Med. J.* 20, 352–353
19 Reidpath, D.D. and Allotey, P.A. (2001) Data sharing in medical research: an empirical investigation. *Bioethics* 15, 125–134
20 Vogeli, C. *et al.* (2006) Data withholding and the next generation of scientists: results of a national survey. *Acad. Med.* 81, 128–136
21 Anderson, N.R. *et al.* (2006) On the persistence of supplementary resources in biomedical publications. *BMC Bioinform.* 7, 260
22 Evangelou, E. *et al.* (2005) Unavailability of online supplementary scientific information from articles published in major journals. *FASEB J.* 19, 1943–1944
23 Kaye, J. *et al.* (2009) Data sharing in genomics – re-shaping scientific practice. *Nat. Rev. Genet.* 10, 331–335
24 Gore, S.M. *et al.* (1977) Misuse of statistical methods: critical assessment of articles in BMJ from January to March 1976. *Br. Med. J.* 1, 85–87
25 Hurlbert, S.H. and White, M.D. (1993) Experiments with freshwater invertebrate zooplanctivores – quality of statistical analyses. *Bull. Marine Sci.* 53, 128–153
26 McGuigan, S.M. (1995) The use of statistics in the *British Journal of Psychiatry*. *Br. J. Psychiatry* 167, 683–688
27 Panel on Scientific Responsibility and the Conduct of Research (1992) *Responsible Science: Ensuring the Integrity of the Research Process* (Vol. I), National Academy Press
28 Allison, D.B. (2009) The antidote to bias in research. *Science* 326, 522–523
29 Bumpus, H.C. (1898) Eleventh lecture. The elimination of the unfit as illustrated by the introduced sparrow, *Passer domesticus*. (A fourth contribution to the study of variation.) *Biol. Lectures: Woods Hole Marine Biological Laboratory* 209–225
30 Lafrenière, D. *et al.* (2009) HST/NICMOS detection of HR 8799 b in 1998. *Astrophys. J.* 694, L148–L152
31 Piwowar, H.A. *et al.* (2007) Sharing detailed research data is associated with increased citation rate. *PLoS ONE* 2, e308