# Do we need an EcoBank? The Ecology of Data-Sharing

Simon A. Queenborough, Ira R. Cooke & Mark P. Schildhauer

Are climate scientists involved in a global conspiracy of epic proportions, intent on driving civilization back to the Middle Ages? Probably not. But the recent overt clash between climate scientists at the University of East Anglia (UEA), Norfolk, UK, and groups sceptical of climate change, sparked by a computer hacker, led to a great deal of publicity and perceived embarrassment for the scientific community (e.g. http://www.realclimate.org). A prime charge of the 'sceptics' is that leading climate scientists do not make much of their data freely available, making it difficult for a broader community to review and validate the alleged patterns in the data. One of the challenges for these scientists is that the climate data are contributed by a large number of disparate researchers, and getting them all to agree on whether and how best to make the data available is tricky. While this incident brought the issue of data sharing into the minds of the general public relative to climate change, the relevance of ecological investigations to many matters of public interest suggests that ecologists will increasingly need to deal with this issue as well

Data sharing has a long tradition, but in the past it could only be done through personal exchange. The recent development of online databases now means that data can be instantly shared with the public and the scientific community. Some disciplines have been much faster to adopt this than others. In particular, near total data sharing via online repositories is now the norm in molecular biology (e.g. Genbank, http://www.ncbi.nlm.nih.gov/). Databases exist for all kinds of biological molecules, experimental protocols, and even raw instrument data, all of which are extensively cross-referenced and designed to be mined by desktop software. Few would doubt that this move towards data sharing has been a huge benefit to molecular biology as a whole, yet many other scientific disciplines, including ecology, have been slow to follow suit.

There are some good, and some not-so-good reasons why ecologists have been slow to embrace an online data sharing world, but first let's list some of the reasons why it's important that we do.

First, sharing data allows greater insights and ideas to be gained from more people looking at the same data. Ideally, this leads to increased understanding of the natural world. Data producers may find collaborators simply by listing data on a repository, and evidence suggests that doing so can dramatically increase citation rate (Piwowar *et al.* 2007). In addition, much time and effort chasing datasets for meta-analyses could be saved.

Second, much ecological data is inherently complementary with other environmental data, making it useful for synthetic or integrative studies. By joining your data with others, you often gain additional 'columns' (covariates for your analyses) in your joint data set, as well as gaining greater statistical power from combining data (greater sample size). Typically, data-sharing also expands your data to cover larger geo-spatial areas or longer time scales, so your analyses provide more general support for a pattern.

Third, it is a crucial element of scientific rigour that studies be repeatable and reproducible. Scientists can check one another's hypotheses by repeating experiments in their own labs, but reanalysis of another scientist's data is often the only option for ecological studies. Most journals stipulate that data must be made available if required, although this seems to be rarely followed through. However, the possibility of allowing other scientists to verify your analyses using your data, and fear of being discovered *in flagrante* committing an egregious error, may increase the rigour of one's own data collection and analysis. Furthermore, the process of putting data into a repository can be useful in its own right since it should enforce thorough and standardised documentation, perhaps even reminding us to note down important meta-data that would otherwise have been forgotten.

Fourth, most scientists working today are funded by public money, and one could argue that the public has a right to view the data that they paid for in taxes. Many ecologists are unaware that the US National Science Foundation, the UK Natural Environment Research Council, and others, oblige data collected under their aegis to be made available a reasonable amount of time following collection. This applies not only to academic institutions, but also to much data collected by government, which is increasingly being made public (http://data.gov.uk).

If sharing data is beneficial to individual data providers and to the science of ecology as a whole, and if journal and funding agency policies all agree that data should be made available, why are so few ecological datasets deposited online? Perhaps it is because the interpretation of ecological data so often

depends upon the way it was collected. Genetic sequences, in contrast, are uniform in structure and to a large extent interpretable outside the data collection context. This not only makes DNA data easy to describe, but makes their usefulness in meta-analyses much more obvious. These factors almost certainly contributed to the great success of Genbank, a central open-access database for DNA data, in which any sequence must be deposited before publication.

The great majority of scientific data, in ecology and in other disciplines, requires considerable contextual metadata to be useful. Recently there has been a push to share this much broader class of data. In ecology, these efforts include many small to medium sized repositories, specifically accepting data of a particular type, or even from a single long-term experiment. Examples include the Amazon Forest Inventory Network, the UK's Biological Records Centre (http://www.brc.ac.uk), the NERC Center for Population Biology's Global Population Dynamics database ( http://www3.imperial.ac.uk/cpb/research/patternsandprocesses/gpdd), the SCAR-MarBIN portal (http://www.scarmarbin.be) and others. While these repositories have been, and will continue to be extremely useful, they fail to provide the kind of 'EcoBank' type database where data of any type could be deposited. Such a generic database is a prerequisite for journals that mandate sharing of data on publication.

Although we can now publish almost any type of ecological data online there is a danger that such data will end up in a fractured state, partly negating the benefits that online sharing brings. The first ecological data archives were no more than giant ftp servers where data files were deposited with no thought to their future use. Data archiving is not the same as data sharing; and archived data is as good as lost if it cannot be reused. Solving this problem isn't necessarily a matter of creating a single monolithic 'EcoBank' type database, but it does require much greater coordination between databases than exists currently. A crucial component of such coordinated data sharing, which has been slow to develop in ecology, is the description of the data that allows future users not only to understand the data variables and collection methods, but also to interpret them in a scientific context. Meta-data is 'data about data'. Describing where data were collected, how and by whom are important aspects that need to be recorded. Perhaps the most comprehensive meta-data schema currently in use within ecology is the Ecological Metadata Language (EML) http://knb.ecoinformatics.org/software/eml/eml-2.1.0/index.html, its purpose being; "To provide the ecological community with an extensible, flexible, meta-data standard for use in data analysis and archiving that will allow automated machine processing, searching and retrieval." The idea is that all data should be described in some minimal but consistent way, with specialist datasets implementing their own extensions of the basic standard. Further to meta-data, development of ontologies that unambiguously define terms can help to describe the detailed semantic content of scientific data. Ontologies are logical, hierarchical and use a constrained vocabulary, much like taxonomic species descriptions, and can be used to create formal specifications for describing data (Jones *et al* 2006).

Unfortunately, very few ecological repositories require data to be entered explicitly according to a published standard like the EML and efforts to create such standards are poorly funded in comparison to those in molecular biology (Madin *et al* 2008). Nevertheless, the Knowledge Network for Biocomplexity, or KNB (http://knb.ecoinformatics.org) represents a successful, distributed data archive based on EML, that is being used by the International Long Term Ecological Research Network (http://www.ilternet.edu) as well as synthesis centers like the U.S. National Center for Ecological Analysis and Synthesis, NCEAS. Alternative data archives such as Dryad (http://www.datadryad.com), are also emerging that serve overlapping communities, yet are based on standards other than EML. This proliferation of alternative data repositories based on incompatible meta-data standards fragments the communities' information resources, such that ecologists wishing to share data will not derive the full benefits that such standards bring, especially relative to high data visibility via automated cross-referencing and searching mechanisms. A large part of the success of molecular biological databases is due to standards. For example the Swissprot protein database is able to cross-reference over a dozen other databases because each of them has its own published standard.

Of course, development of ecological data standards isn't easy, it will require us to think hard about what information to record, and importantly, to agree on and publish what we come up with. However, the reward of having standards, as well as the beneficial process of discussion that leads to their publication is something we should not miss out on. Standardization will also require that various existing repositories, such as the KNB and Dryad, strive for strong compatibility between their standards, by developing "meta-standards". These types of 'data/meta-data' confederation efforts are being addressed through projects such as the NSF-funded DataONE initiative (http://www.dataone.org)

While technological hurdles such as lack of standards are important, negative perceptions about the consequences of data sharing may be just as much of a barrier to data-sharing. Many ecologists have deeply intimate associations with their data, having collected it over many years, often in adverse physical and psychological conditions. Moreover, ecological data often comprise observations that could only be collected by someone with advanced training or experience, for example being able to rapidly and accurately identify taxonomic identities of organisms, or being able to discern subtle behaviours or even physical features within the natural environment that would not be evident to a more naïve observer (e.g. evidence of plant-grazing, change in vegetation structure, etc.). Data is hard-won and is hard work to part with. It is 'my' data and should remain so, especially if I intend to keep publishing papers from it! In an effort to ease this potential cause of suffering, and recognise the scientific expertise and creativity often involved in simply collecting data, the Ecological Society of America now publishes citable Data Papers that are peer-reviewed with the abstracts published in the appropriate print journal (www.esapubs.org. archive). To 2008, 1179 Data Papers had been published.

Fear of being scooped by a competitor, or indeed an evil referee, looms large on the list of negatives. Many journals in molecular biology prevent referees from scooping authors by allowing data to be withheld until publication. Ecological journals could go further if necessary, providing a longer period of exclusivity. We may yet see the day when all ecological journals insist upon a data archive accession number prior to publication. Nature Publishing Group journals have already adopted such a policy, and the American Naturalist will soon do so (Whitlock *et al.* 2010).

Another objection often made to sharing data is that other researchers need to know the data intimately in order to analyze it properly. This returns to the issue of meta-data standards as well as the need for ontologies, which if sufficiently detailed could enforce high quality documentation and potential for accurate interpretation and re-use. Even without standards however, high quality documentation can be provided in a simple README file along with the data. Providing this file ensures that the data is much less likely to be mis-analyzed, even by the original data collector (who might forget details over time). A README, however, will not be as amenable to computerized processing, such that investigators might face the arduous task of wading through an overwhelming number of READMEs, to find data sets containing measurements of interest to them.

Finally, there is a fear that data depositors will be ripped off by opportunistic analysts. If just about anybody can get their sceptical hands on your data and produce a dubious Excel graph, how should we trust that the resulting analyses have not been corrupted or manipulated, and that the original data collectors are properly acknowledged? Actually, evidence suggests that papers that publish data online are cited more, indicating that proper acknowledgement is the norm (Piwowar *et al.* 2007). In addition, data that are shared will grow in importance due to the greater number of researchers who use them. Even if a small percentage of these users fail to properly acknowledge the data collector, the net effect will almost certainly be an increase in scientific standing via citations and collaborations.

Fundamentally, there is currently very little professional incentive for ecologists to share data online. We are not rewarded for doing so, data is often not citable, and its position seems to be relegated below that of teaching and professional service. For all the societal, professional and personal advantages of data-sharing described above to be fully realised, data, like publications, must become part of the hard currency of science.

Trust is fundamental in human endeavour, and especially so in science. The technological difficulties of dealing with disparate sources and formats of data are being overcome. We are now able to share data easily with scientists and the public all over the world. Although trusting that others will not abuse that data remains a concern for many scientists, proven mechanisms exist to deal with these issues: for example, data remaining effectively copyrighted for a number of years post-experiment, and mandatory archiving of data with proper accompanying documentation. Would an open-access database of all the climate data silence the sceptics? Highly unlikely. But having such a verified, official database would enable any scientist, professional or otherwise, to check the data for themselves and have allowed the 'debate' to be about science rather than skullduggery.

Feel free to contact any of us for discussion of this issue; more technical comments or queries should be addressed to Mark Schildhauer (**schild@nceas.ucsb.edu**).

## REFERENCES

Jones, M.B., Schildhauer, M.P., Reichman, O.J. & Bowers, S. 2006 **The new bioinformatics: Integrating ecological data from the gene to the biosphere. *AREES* 37, 519-544.**

Madin J.S., Bowers, S., Schildhauer, M.P. & Jones, M.B. 2008 **Advancing ecological research with ontologies. *TREE* 23, 159-168.**

Piwowar, H.A., Day, R.S. & Fridsma, D.B. 2007 **Sharing detailed research data is associated with increased citation rate. *PLoS ONE* 2, e308.**

Whitlock, M.C., McPeek, M.A., Rausher, M.D., Rieseberg, L. & Moore, A.J. 2010 **Data Archiving. *Am. Nat.* 175, 1-2.**

Simon Queenborough and Mark Schildhauer are at the National Center for Ecological Analysis and Synthesis in Santa Barbara, California; Ira Cooke is in the Department of Biochemistry at La Trobe University in Melbourne, Australia.

**The BES Publications team are currently seeking the opinion of all BES members on this issue: please see p00**