Global Change Biology

Global Change Biology (2012) **18**, 2102–2110, doi: 10.1111/j.1365-2486.2012.02693.x

REVIEW

Advances in global change research require open science by individual researchers

ELIZABETH M. WOLKOVICH*†, JAMES REGETZ‡ and MARY I. O'CONNOR†

*Department of Biological Sciences, University of California, San Diego, 9500 Gilman Drive #0116, La Jolla, CA 92093-0116, USA, †Department of Zoology, University of British Columbia, 6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada, ‡National Center for Ecological Analysis and Synthesis, 735 State Street, Suite 300 Santa Barbara, CA 93101, USA

Abstract

Understanding how species and ecosystems respond to climate change requires spatially and temporally rich data for a diverse set of species and habitats, combined with models that test and predict responses. Yet current study is hampered by the long-known problems of inadequate management of data and insufficient description of analytical procedures, especially in the field of ecology. Despite recent institutional incentives to share data and new data archiving infrastructure, many ecologists do not archive and publish their data and code. Given current rapid rates of global change, the consequences of this are extreme: because an ecological dataset collected at a certain place and time represents an irreproducible set of observations, ecologists doing local, independent research possess, in their file cabinets and spreadsheets, a wealth of information about the natural world and how it is changing. Although large-scale initiatives will increasingly enable and reward open science, we believe that change demands action and personal commitment by individuals – from students and PIs. Herein, we outline the major benefits of sharing data and analytical procedures in the context of global change ecology, and provide guidelines for overcoming common obstacles and concerns. If individual scientists and laboratories can embrace a culture of archiving and sharing we can accelerate the pace of the scientific method and redefine how local science can most robustly scale up to globally relevant questions.

Keywords: code management, data management, global change ecology, open science, scientific method

Received 19 December 2011; revised version received 22 February 2012 and accepted 24 February 2012

Introduction

In 1953, a group of pediatric cancer specialists met to discuss the process and progress of scientific research in their field (Unguru, 2011). At the time, the survival rate for the most common childhood cancer, a type of leukemia, was <4%. The result of the meeting was a transformation of cancer research in the United States: from small scale, local laboratory work to team science and cooperative research. In 1954 multi-institutional studies combined with strong data sharing policies began uniting the research efforts of 40 hospitals (Unguru, 2011), and the cure rate started to climb. Today, the survival rate for the same childhood cancer is 94% and the field worldwide recognizes cooperative, collaborative science as a driving force behind this change (Devidas *et al*., 2010).

In the same year, cooperative cancer research began, botanist R. S. R. Fitter began to collect what would become a 47-year record of flowering times in his local area (Fitter & Fitter, 2002). Today, Fitter's dataset represents one of the most important records of how plant species have shifted with climate change: researchers have used it further to show how climate responses are linked to trait conservatism and species distributions (Davis *et al*., 2010; Hulme, 2011). Such wide use is in no small part due to the fact that Fitter and his co-author took the unusual step of publishing the dataset with their article in *Science* (Fitter & Fitter, 2002), freeing others to use the data.

This example is inspiring. Unfortunately, it is outside the norm for global change ecology, where – compared with other fields working toward data sharing (Schofield *et al*., 2009; Igo-Kemenes, 2011; Overpeck *et al*., 2011) – progress has been unusually slow. Despite over 15 years of focused efforts toward developing infrastructure for sharing data, in community ecology and global change ecology participation by researchers today remains the extreme exception (Michener *et al*., 1997; Fegraus *et al*., 2005; Parr & Cummings, 2005; Nelson, 2009; Reichman *et al*., 2011). Yet anthropogenic environmental change makes most ecological field data collected effectively irreproducible, critical snapshots in

Correspondence: Elizabeth M. Wolkovich, tel. + 604 822 0862; fax + 604 827 5350, e-mail: wolkovich@biodiversity.ubc.ca

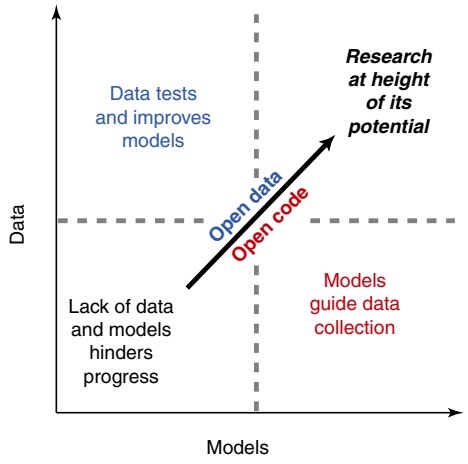

Figure 1 Accelerating research on applied issues requires high availability of both data and models. Adapted from Holling (1978).

time that cannot simply be remeasured. Ecologists now realize studying 'Nature' – a pristine entity – is an impossibility (Sih *et al.*, 2004). Thus, every dataset has potential value for synthetic projects that explore ecological change in space and time.

Rapid and major advances in global change research are supported by high data and model availability (Fig. 1), which requires a shift toward open science at the individual level. Herein, we argue that individual scientists can have a tremendous, positive impact on the pace of global change science by archiving and sharing their data along with any relevant data processing steps and analytical procedures (hereafter referred to simply as 'code' in reference to computer source code, although reproducible steps can be captured in non-code forms as well). First, we provide a brief overview of how local, ecological data is used in global change science. We then argue that the top-down initiatives for data- and code sharing already underway (Mervis, 2010; Whitlock *et al.*, 2010) are necessary, but not sufficient: individual scientists must actively participate in open science for it to succeed. We outline simple steps that individuals can take in their careers to overcome common obstacles and develop the necessary skills to archive and share data and code, with the goal of moving data and code from the individual to public level. Such a shift will benefit individual ecologists and research groups, as well as whole fields of research, with cascading effects for the contribution of ecological research in the sciences and in ecosystems across the globe.

Poor datasharing slows global change research

Across fields, papers and initiatives building on open data access have led to breakthroughs and critical insights (Piwowar *et al.*, 2008; Carpenter *et al.*, 2009). In medicine, collaborative team science, including shared data and resources, has led to large improvements in cancer treatments (Reaman, 2002) and in the early diagnosis of Alzheimer's disease (Kolata, 2010; August 12). In ecology, shared data has underlied recent advances in understanding global ecosystem services, such as decomposition (Wall *et al.*, 2008) and pollination networks (Rezende *et al.*, 2007). In ocean sciences, the invention and advancement of large-scale plankton monitoring, begun by one scientist, has documented major shifts in ocean food webs and algal blooms over recent decades (Edwards & Richardson, 2004; Edwards *et al.*, 2010). To date, the data – which are publicly available – have resulted in over 1300 publications (Sir Alister Hardy Foundation for Ocean Science, 2012).

In global change research, synthesis efforts have been critical to showing how climate change impacts ecological systems (Root *et al.*, 2003; Rosenzweig *et al.*, 2008), with long-term time-series data (Hegerl *et al.*, 2010) acting as the standard for detecting and attributing biological shifts to climate change. Yet very few of these datasets are available, limiting our ability to answer even basic questions, such as how frequently climate change impacts are actually observed (Parmesan & Yohe, 2003). Further, the iterative process of testing and fitting models – conceptual, theoretical or simulation – to data can advance only as fast as data availability and model development time allow (Fig. 1). We argue that global change research, especially in the field of ecology, is stymied by inefficient data use and low-model availability, hindering research applications. Alongside an urgent need to collect new data to answer global change questions, advances can be made by synthesizing existing data to document historical change and baseline conditions.

Despite recent progress at the level of institutions and the development of new infrastructure for data archiving (Reichman *et al.*, 2011), access to ecological data is still extremely low (Nelson, 2009), resulting in a grossly impeded pace of research as synthesis projects require extensive emails and extraction of data by hand from publications. In addition to being time intensive, data extraction from publications usually yields only derived data (e.g., treatment means and standard errors), which confines researchers to a narrow set of meta-analytic tools that handle such data types (Cooper *et al.*, 2009). In contrast hierarchical models based on raw data allow more robust tests and provide not only an understanding of effect sizes but also comprehensive information on the error structure of the data. These methods, thus, can provide novel insights and improve future experimental designs; not surprisingly, much research in the medical sciences has already

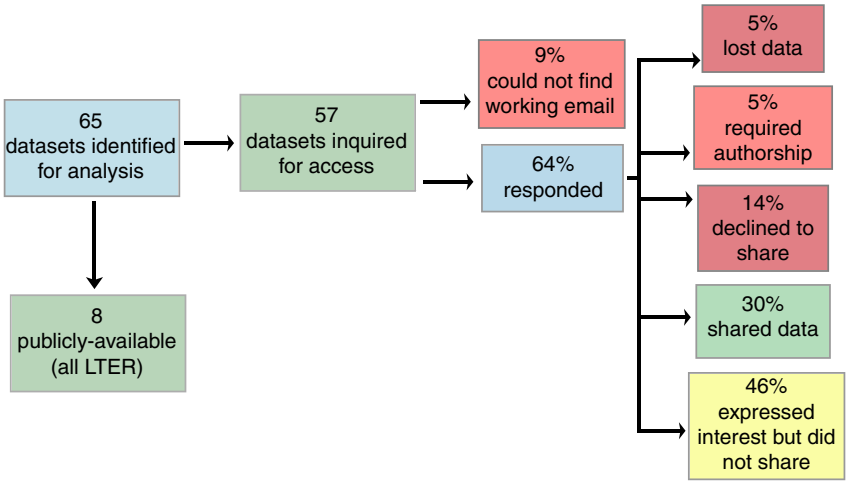

Figure 2 For a recent meta-analysis on the effects of temperature on plant phenology one author (EMW) identified 65 datasets of interest (a nearly equal mix of observational long-term data and experimental results) and attempted to obtain the raw data for them. For eight datasets data were already publicly available online (all through the US Long-Term Ecological Research program), for the remaining 57 the author sent over a 100 emails to request data: 64% of emails received a response, 27% did not (for 9% of emails multiple email addresses failed). Of those responding, 30% shared data.

switched to this method for meta-analyses (e.g., Alberti *et al.*, 1995; Concato *et al.*, 2000; Baigent *et al.*, 2005).

In global change ecology, however, raw data are rarely available. For a recent meta-analysis one of the authors (EMW) attempted to track down raw experimental and long-term monitoring datasets (Fig. 2). After >100 emails inquiring for nearly 60 datasets over 10 months, only 11 datasets were eventually shared. Yet well over half of those who responded expressed interest in sharing their data.

What prevented the large group of researchers who expressed interest in sharing from following through? Although we do not have information on this sample, we can speculate that the reasons are similar to those addressed in several recent reviews of perceptions of the risks of data sharing, and the realities (Table 1, see also Parr & Cummings, 2005; Smith, 2009; Overpeck *et al.*, 2011; Tenopir *et al.*, 2011; Whitlock, 2011). Such study suggests many common concerns may be based more on fears than reality and that sharing can have many benefits for the sharer. For example, papers accompanied by shared data are cited more often (Piwowar *et al.*, 2007; Zimmerman, 2008). Further, many common issues, such as attribution and credit for data, are already being tackled by organizations and initiatives focused on promoting and providing a means to share data (Tables 1 and 2).

More persnickety issues, however, remain. Open science means increased collaboration at numerous stages in the process – data sharing, code sharing and idea sharing – and concerns have been raised over whether such sharing and collaboration are properly credited by traditional citation metrics and methods (Raff, 2003; Kueffer *et al.*, 2011). Properly crediting data is a real issue, and one that – given more synthetic and large-scale research – will need to be addressed with cultural and institutional changes. Ideally, researchers repurposing data will involve data holders throughout the project – from hypothesis refinement to publication – since data holders usually have additional insights and understanding of their data that is difficult to capture in metadata. How to involve data holders on very data-rich projects (i.e., those projects that use tens or hundreds of individual datasets) requires additional consideration. Authorship-in-exchange-for-data policies may discourage many researchers from using all available datasets. In addition, such policies have become problematic in cases of blame-assignment when issues with a study arise (Kennedy, 2003). Further issues could arise if data-authors do not agree with findings and thus pull data from a particular study. Although some have argued for inclusive authorship policies (Kueffer *et al.*, 2011), many guidelines suggest data alone does not yield authorship (Weltzin *et al.*, 2006). More large-scale, data-reuse projects in global change ecology are probably required before community agreement on new policies will be possible. Eventual guidelines, however, will have to balance the demands of crediting the valuable resource of data, while not hindering research. At this point, we believe authorship considerations must be handled on a case-by-case basis.

Another concern is whether data sharing creates a system where 'cheaters' – those who do not share data, but copiously use public data – will reap the greatest benefits. Although this is possible according to basic

Table 1 Common concerns associated with sharing data and code and responses to those concerns

Perceived concern	Reality
Insufficient attribution and credit	Increased citation of paper with data; publication credit for data (e.g., *Ecological Archives*, Dryad DOIs)
Uneven playing field between those who share data and those who only *use* shared data	Promotes equality on some scales (international researchers, those without funding to collect data, etc.); first-hand knowledge of data will remain invaluable
Risk of revealing sensitive information (e.g., harvest data, rare species locations)	Many options available to scramble information
Lost opportunity to publish on one's own data	No need to post all long-term data; e.g., with 30 years of data, publishing years 1–10 makes it easier for you (or someone else) to post the rest at a later date (Roberts, 2002)
High financial costs	Overall costs are quite low (Piwowar *et al*., 2011), but funding agencies should consider funding data management costs in proposals
High personal time costs	There are a variety of options now to post data – some require minimal time, although greater investment yields greater utility to you and others (Borer *et al*., 2009)
Potential for data misuse	Rarely – if ever – reported, further, careful metadata should mitigate this (Whitlock, 2011)
Fear of exposing imperfect coding practices	'Publish your code: it is good enough' (Barnes, 2010)
No mechanism for non-programmers to share analyses	Visual workflow software: Kepler, Taverna, and VisTrails (see also Table 2)
Concern about longevity of repositories	Even with uncertain long-term futures, initiatives like DataONE will certainly increase lifespan of data relative to ad hoc individual data archiving practices
No personal benefit to sharing	Using repositories relieves you of the hassle of managing/archiving your own files; access your content from anywhere; emerging automated tools will allow further benefits, e.g., automated taxa scrubbing, easy integration with other online datasets

game-theory, in reality the outcome of open science is more likely to involve mutually beneficial partnerships, rather than parasitism. Most ecologists who work with existing data recognize the value of collaborating closely with those who are directly involved in the data collection. The exceptions typically involve broad data synthesis and re-analysis activities for which the specific research questions are well beyond the scope of what could be answered by any individual data collector. Moreover, those who use existing data generally also make their own data available (Zimmerman, 2008). At a more global level, shared data can grossly benefit scientists from under-represented countries, providing them with access to large datasets they could never develop empirically given local funding levels. At the level of the scientific method, data synthesis does not slow data collection but rather, guides and encourages it (Fig. 1).

Recent events, such as those surrounding stolen emails from the University of East Anglia, have made it clear that improved descriptions of data manipulation and analysis steps must be a part of this new paradigm of open science. Increasingly, global change science is indirect (Mesirov, 2010), that is, datasets are too large to visually handle and thus researchers must process data without seeing every row personally. Revealing one's potentially imperfect and inelegant computer code and related details may be uncomfortable at first, but the community must accept that computational decisions should be subject to review and refinement much like the rest of our science (Ince *et al*., 2012). With greater openness, the opportunity to more rapidly converge on better analytical methods is immense (Raymond, 1999). As with data, the research community will benefit from increased opportunity to reuse and extend shared code, or to collaborate in enhancing the features of community models. Expectations of shared code, as well as data, also level the playing field of shared resources: those re-purposing data often have invested large amounts of time in developing and testing sophisticated code, from which others, including data collectors, could benefit. Further, openness with analytical code would make model comparison a less-daunting task, and would engage researchers who may not be interested in developing models from scratch, but who could improve or apply such models. Code sharing would also allow routine tests of one hypothesis across multiple systems and would remove some of the roadblocks to interdisciplinary work. Ultimately, it would fundamentally shift the pace of global change research, aiding progress toward science that operates at its highest potential (Fig. 1).

Solution: individual commitment to sharing at the grassroots level

Open science will not progress until researchers recognize that every dataset can make multiple contributions

Table 2 Actions and related skills needed for data and code management and sharing. Because many articles and online sites are devoted to this topic we review skills and resources here briefly

Action	Related skills	Advantages	Example tools/services
Archive and share data in-house (Step 1)	Data management, metadata creation (Borer *et al*., 2009)	Data preserved for individual researcher to reuse and share with colleagues and students	Data management desktop software (Morpho[1]); metadata standards (EML[2], FGDC CSDGM[3], Darwin Core[4]);
Archive and share code in-house (Step 1)	Workflow management, from the field/ laboratory to analysis (Borer *et al*., 2009)	Analyses archived for individual researcher to re-run, reuse, and share with colleagues and students	Scripted languages (R[5], MATLAB[6], Python[7]); version control software (subversion[8], Git[9]); scientific workflow software (Kepler[10], Taverna[11])
Archive and share data publicly (Step 2).	Data management, metadata creation	Data can be used to answer unexpected questions; satisfy funder or journal data policies; redundant backup	DateONE[12] member nodes (KNB, Dryad[13], Mercury[14]); PANGAEA[15], NCDC[16]
Archive and share code publicly (Step 2)	Workflow management, from the field/ laboratory to repository	Analyses/models can be reused by others; definitive statement of what you did; opportunity for feedback, improvement, collaboration	Data repositories that accommodate code (KNB); workflow repositories (myExperiment[17]); code hosting services (GitHub[18])
Use and cite existing public data when possible (Step 3)	Awareness of relevant data repositories; ability to use relevant data query interfaces and tools	Preliminary data to test and refine hypotheses available to all; encourages beneficial collaboration, synthesis, and sharing	DataONE, GBIF[19], NASA GCMD[20]
Include data and code management skills when teaching and mentoring (Steps 1 and 3)	Data management, workflow management (Borer *et al*., 2009; Whitlock, 2011)	Promote a new culture of openness	DataONEpedia[21], Software Carpentry[22]

[1]http://knb.ecoinformatics.org.
[2]http://knb.ecoinformatics.org/software/eml.
[3]http://www.fgdc.gov/metadata/csdgm.
[4]http://www.tdwg.org/activities/darwincore.
[5]http://www.r-project.org.
[6]http://www.mathworks.com/products/matlab/index.html.
[7]http://python.org.
[8]http://subversion.apache.org.
[9]http://git-scm.com.
[10]http://www.kepler-project.org.
[11]http://www.taverna.org.uk.
[12]https://www.dataone.org.
[13]http://datadryad.org.
[14]http://mercury.ornl.gov/ornldaac.
[15]http://www.pangaea.de.
[16]http://www.ncdc.noaa.gov/paleo/.
[17]http://myexperiment.org.
[18]http://myexperiment.org.
[19]http://data.gbif.org.
[20]http://gcmd.gsfc.nasa.gov.
[21]https://www.dataone.org/dataonepedia.
[22]http://software-carpentry.org.

to scientific understanding. First, the dataset can serve the original intent of the researcher to address an immediate or local question or hypothesis. Second, the dataset represents a spatial, temporal or taxonomic replicate potentially useful in future synthetic projects likely not anticipated at the time of original data

collection. For example, long-term data and re-surveys of species' phenologies and distributions – collected originally for a variety of unrelated purposes – have underlied most evidence of biotic responses to climate change (Parmesan *et al*., 1999; Root *et al*., 2005; Walther *et al*., 2005). Similar extensions of data have allowed global assessments of marine ecosystems (Halpern *et al*., 2008) and a diversity of paleoecological studies have shown how species' ranges shift with climate (Jackson *et al*., 1997; Gray *et al*., 2006). These unanticipated uses often manifest in collaborative research years after the data were originally collected. With these ultimate uses in mind, traditional methods of data storage and curation are outdated and often impede achieving the second goal.

Although many global change ecologists may see the merits and have interest in data sharing (Fig. 2), the apparent lack of rewards for individuals and a rampant lack of training in good practices likely limit most. While initiatives from funding agencies, societies and journals are underway (Stein, 2008; Mervis, 2010; Whitlock *et al*., 2010) and may have great power in altering the reward structure, we believe individuals have the greatest power for change. Further, the resources for individuals to effectively manage data and code are already in place. Herein, we describe a three-step approach that outlines how individual researchers can identify and develop the skills needed and actions they can take now to promote data and code sharing in their community. We note that commitment toward better management is the main needed action; exact resources to use and how much training toward the most advanced data and code management researchers decide to pursue will vary (Table 2).

Step 1. Archive and share data and code in-house

Managing data and code so that it can be useful to researchers months and years later can seem painful and tedious when done as an afterthought. However, by considering the initial use and longer term stewardship of data and code early in a project researchers can establish a foundation for effective management and in-house sharing.

Researchers should consider their data's life cycle (Lee *et al*., 2009; Strasser *et al*., 2011) – how will data collected in the field or laboratory and its documentation (metadata) be organized initially and in computer format? How will it be backed-up and where will it be stored eventually? Most researchers accomplish elements of this process easily as part of their scientific process. For many, the needed advance is to provide documentation and long-term storage. Excellent metadata, for example, usually requires centralizing only a few key pieces of information: when, where, and how data were collected, and by whom. Such information generally takes only a few minutes to include if done alongside data entry, but because it is not a process ecologists are generally trained in, it is often skipped. This first step, however, of providing basic metadata overcomes a major challenge to future effective use of the data. The need for good data curation has long been recognized and steps toward this goal nicely outlined by Borer *et al*. (2009). While sometimes seen as tedious, including more detailed metadata about the the physical format and structure of the data (e.g., number of columns and rows, data types and units) ensures that future downstream analyses can be performed with less risk of misinterpretation.

Researchers also need to move toward more deliberate consideration of their planned analyses, spanning all steps of an analytical workflow that transforms raw input data to final results. What tests will be run on what subsets of the data? What program will be used? At the outset, mentally conceiving of this process in the form of a flow diagram can yield a more coherent plan for how best to organize data and, with practice, can streamline and accelerate analyses (Goodchild, 2000). Particularly for synthetic projects where data are transformed or incorporated into a summary statistic, documentation of the flow from data to manuscript figures and models is invaluable (Ellison, 2010). Such documentation not only ensures the reproducibility of scientific work but also allows others to build on the work by expanding datasets or adapting analyses to different situations and research questions. Documentation of analytical workflows can take several forms. Scripted languages (such as R, MATLAB, or Python) are ideal because they allow for a written file that, when run, can read in the data and carry out all relevant computations – including creation of figures and tables – without error-prone manual intervention. This approach is advantageous because no additional documentation is required and it has direct benefits to the individual: complex code components can be reused, and manuscript revisions are far easier when the analytical process can be re-run identically and rapidly. A complete history of script modifications, and the final versions themselves, can be easily managed using version control software that tracks, dates, and documents changes to the analysis process (Table 2). Alternatives to scripting languages include workflow software, which is freely available to help track both data management and analytical workflows (Table 2).

Following these practices in a research group or a particular research program has clear benefits. First, researchers at different career stages can learn together how to use software, to create useful metadata, and

archive details of data collection and analysis properly. Research groups can discuss issues in developing their data- and code-curating procedures, and by engaging members of a group with particular strengths, all members can learn what works well. Finally, research groups often perform research on one or several related themes, with students working on similar species, questions or sites. The potential for synthetic projects within research groups is great, if data and code from past students and postdocs are clearly archived and available to future students. The training and research opportunities for students and PIs associated with good data and code-curating practices are clear incentives to spend some time on this issue in laboratory meetings and student training.

Step 2. Archive and share data and code publicly

Once good practices are developed within a research group, making data public is actually quite simple. Thus, researchers who adeptly manage data and code in-house are well poised to make data public, as required by many new grant agencies' and journals' requirements (Mervis, 2010; Whitlock *et al*., 2010). Data can be made public at different levels: full datasets can be archived with journals or data repositories (Table 2). Or, metadata alone (without the full dataset) can be posted to some repositories (e.g., KNB), leaving distribution of the full dataset under the control of the researcher. Posting metadata alone at an early stage makes posting data when it is finally available a less-daunting task. Researchers should, however, ideally post raw data and make them freely available with useful metadata and accompanying processing and analysis scripts; if posting raw data is not possible due to collaborator restrictions, researchers should supply derived data products (e.g., means, sample sizes, and standard errors). In all cases, laboratory groups should create metadata in a format endorsed by community-sanctioned repositories (e.g., KNB, Dryad, and Mercury) so posting data is simple and direct.

Step 3. Engage in a culture of open science

Individual researchers currently have myriad opportunities to promote and influence the process of data- and code sharing, and associated collaborations, in global change ecology. They can simply engage in discussions on data sharing to help others see the potential to advance science if data and code were managed and shared more thoughtfully. They can review data management plans in proposals carefully, pointing out issues and suggesting improvements whenever possible. We especially believe including data and code management skills as a part of science when teaching and mentoring can help shift the practice of science. This is a particularly important option for funded projects where teaching is an integral goal (such as the US National Science Foundation's IGERT program and training initiatives at large synthesis centers).

Looking forward: progress toward a new mode of science

Improving success of top-down initiatives through grassroots support

Increased recognition of the possibilities data and code sharing provide (Fig. 1) can, in turn, create greater opportunities for successful larger data and code initiatives. Other fields have shown that researcher support and request for top-down actions in data sharing can create sweeping changes in the practices of individual researchers. Journals require genetic data to be rapidly and freely available in central repositories (Field *et al*., 2009; Schofield *et al*., 2009) and we suggest ecological and global change journals should follow the recent suite of evolution journals (Whitlock *et al*., 2010) that now require data related to articles to be posted to public repositories (Savage & Vickers, 2009) and encourage analytical workflow posting as well. Societies and agencies can support such efforts by providing standards of practice for data (file types, suggested repositories) and code (scripting practices and how to capture analytical workflows in non-scripted computer programs) and by offering training sessions or workshops – for example in association with professional societies' annual meetings.

We believe funding agencies have great power to enact change. The European Commission's major funding program, FP7, requires data management plans of proposals and Germany's Research Foundation (Deutsche Forschungsgemeinschaft, DFG) now requests plans for facilitating reuse of data, with certain disciplines requiring formal data management plans. In the US, the National Science Foundation (NSF) recently required a data management plan for all research proposals. We hope these new requirements signal a changing tide in the support and credit given for data in the sciences. Assessment of such data management plans should make usable (i.e., provided with complete metadata), public data a priority. Researchers should suggest how they will manage data and metadata from the field to the public repository, and include information about the total amount of data expected to be archived and when. In the future, we hope national and international funding agencies will review results from previous support with an eye toward both the quality of public data – as well as the publications – produced.

In the US, reviewers of NSF proposals now have the ability to think about, critically review, and reward excellent data management practices of their peers.

Scaling up research: new skills and roles

We have attempted here to outline the argument for and path toward data and code sharing by individual global change scientists: starting a grassroots effort to promote a new perspective and standard for how we value and use global change ecology data in the scientific process (Fig. 1). We believe our above action items implemented in just a small fraction of global change laboratories could rapidly advance our understanding of fundamental issues in global change research. Thereupon, future initiatives can build toward a new mode of science, one that can tackle questions on new scales and deliver verifiable research for policymakers.

As data and code become more widely available, scaling up the scientific process may help reshape attribution. Global change ecologists are often adept at working across thematic scales – from populations to communities, for example – highlighting a skill in developing innovative ideas and concepts. Yet many ecologists also possess skills at well-designed data collection and creative analysis. Currently, however, these skills are not given adequate recognition even though they are critical to first rate science (Fig. 1). In the future, we hope the field of global change ecology will find avenues to give suitable credit to those whose specialized skills have advanced science by fostering collaboration, cultivating valuable intradisciplinary and interdisciplinary teams or by providing high quality, reusable data and analyses. We suspect such a cultural transformation is underway, but we hope our proposal to shift how global change biologists approach data- and code sharing will broaden and accelerate our field's thinking about the best integration of skills, data, credit and rewards.

Conclusions: the future of data in global change ecology

Sharing of data and code does not imply that new data collection will ever become less useful or necessary, but rather that data and code sharing can both advance how useful and insightful new data collection and models are. Ecology today is still most often carried out at local scales ('boots and buckets'), and such study continues to generate observations and information fundamental to documenting and understanding change from the individual organism to the community and ecosystem levels. Insights from this level of science need to be better integrated with insights from large-scale database and model efforts, and conversely, large-scale patterns must be validated and informed by detailed local processes as well.

Today global change ecology has an opportunity for transition. While small-scale local research continues to be critical, projects are increasingly conducted less on local scales (Stokstad, 2011) and more at the regional, national, and global level (e.g., Nutrient Network, NEON). At the same time, climate-induced shifts in organisms worldwide have enhanced public interest in ecological research. This is in large part because of the enormous implications for biodiversity and ecosystem services that in turn can impact basic human needs such as access to water or food. The increasing relevance of research in global change ecology, along with developments in infrastructure that enable distributed storage of, access to, and analysis of data, provide tremendous opportunity to transform how ecologists 'do science' – with greater openness benefiting individuals and the field as a whole.

Acknowledgements

The authors would like to thank J. M. Wolkovich for initial information on data sharing in oncological research, M. Jones, B. Leinfelder, M. Schildhauer, M. Whitlock and three anonymous referees whose comments on earlier drafts improved this manuscript. This work was conducted while EMW was an NSF Postdoctoral Fellow in Biology (Grant DBI-0905806), and also while she was supported by the NSERC CREATE training program in biodiversity research; JR was supported by NCEAS, a Center funded by NSF (Grant EF-0553768), the University of California, Santa Barbara, and the State of California.

References

Alberti W, Anderson G, Bartolucci A *et al.* (1995) Chemotherapy in non-small-cell lung-cancer – a metaanalysis using updated data on individual patients from 52 randomized clinical-trials. *British Medical Journal*, **311**, 899–909.

Baigent C, Keech A, Kearney PM *et al.* (2005) Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90 056 participants in 14 randomised trials of statins. *Lancet*, **366**, 1267–1278.

Barnes N (2010) Publish your computer publish your code: it is good enough. *Nature*, **467**, 753.

Borer ET, Seabloom EW, Jones MB, M S (2009) Some simple guidelines for effective data management. *Bulletin of the Ecological Society of America*, **90**, 205–214.

Carpenter SR, Armbrust EV, Arzberger PW *et al.* (2009) Accelerate synthesis in ecology and environmental sciences. *BioScience*, **59**, 699–701.

Concato J, Shah N, Horwitz RI (2000) Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine*, **342**, 1887–1892.

Cooper H, Hedges LV, Valentine JC (2009) *The Handbook of Research Synthesis and Meta-Analysis* (2nd edn). The Russell Sage Foundation, New York.

Davis CC, Willis CG, Primack RB, Miller-Rushing AJ (2010) The importance of phylogeny to the study of phenological response to global climate change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **365**, 3201–3213.

Devidas M, London WB, Anderson JR (2010) The use of central laboratories and remote electronic data capture to risk-adjust therapy for pediatric acute lymphoblastic leukemia and neuroblastoma. *Seminars in Oncology*, **37**, 53–59.

Edwards M, Richardson AJ (2004) Impact of climate change on marine pelagic phenology and trophic mismatch. *Nature*, **430**, 881–884.

Edwards M, Beaugrand G, Hays GC, Koslow JA, Richardson AJ (2010) Multi-decadal oceanic ecological datasets and their application in marine policy and management. *Trends in Ecology and Evolution*, **25**, 602–610.

Ellison AM (2010) Repeatability and transparency in ecological research. *Ecology*, **91**, 2536–2539.

Fegraus EH, Andelman S, Jones MBMS (2005) Maximizing the value of ecological data with structured metadata: an introduction to ecological metadata language (eml) and principles for metadata creation. *Bulletin of the Ecological Society of America*, **86**, 158–168.

Field D, Sansone SA, Collis A *et al.* (2009) Megascience ,omics data sharing. *Science*, **326**, 234–236.

Fitter AH, Fitter RSR (2002) Rapid changes in flowering time in British plants. *Science*, **296**, 1689–1691.

Goodchild MF (2000) *Communicating the Results of Accuracy Assessment: Metadata, Digital Libraries and Assessing Fitness for Use*, pp. 3–16. Ann Arbor Press, Chelsea, MI.

Gray ST, Betancourt JL, Jackson ST, Eddy RG (2006) Role of multidecadal climate variability in a range extension of pinyon pine. *Ecology*, **87**, 1124–1130.

Halpern BS, Walbridge S, Selkoe KA *et al.* (2008) A global map of human impact on marine ecosystems. *Science*, **319**, 948–952.

Hegerl G, Hoegh-Guldberg O, Casassa G *et al.* (2010) Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Detection and Attribution of Anthropogenic Climate Change, chap. Good Practice Guidance Paper on Detection and Attribution Related to Anthropogenic Climate Change. IPCC Working Group I Technical Support Unit, University of Bern, Bern, Switzerland.

Holling CS (1978) *Adaptive Environmental Assessment and Management*, pp. 67–69. International Series on Applied Systems Analysys, John Wiley & Sons, New York.

Hulme PE (2011) Contrasting impacts of climate-driven flowering phenology on changes in alien and native plant species distributions. *New Phytologist*, **189**, 272–281.

Igo-Kemenes P (2011) *Keeping Data Alive for Long-term Re-use*, pp. 14–15. Alliance for Permanent Access, The Hague, the Netherlands.

Ince DC, Hatton L, Graham-Cumming J (2012) The case for open computer programs. *Nature*, **482**, 485–488.

Jackson ST, Overpeck JT, Webb T, Keattch SE, Anderson KH (1997) Mapped plant-macrofossil and pollen records of late quaternary vegetation change in eastern North America. *Quaternary Science Reviews*, **16**, 1–70.

Kennedy D (2003) Multiple authors, multiple problems. *Science*, **301**, 733.

Kolata G. Rare sharing of data leads to progress on Alzheimer's. New York Times, 12 August 2010.

Kueffer C, Niinemets U, Drenovsky RE *et al.* (2011) Fame, glory and neglect in meta-analyses. *Trends in Ecology and Evolution*, **26**, 493–494.

Lee JW, Zhang JT, Zimmerman AS, Lucia A (2009) Datanet: an emerging cyberinfrastructure for sharing, reusing and preserving digital data for scientific discovery and learning. *Aiche Journal*, **55**, 2757–2764.

Mervis J (2010) NSF to ask every grant applicant for data management plan. *Science* Available at: http://news.sciencemag.org/scienceinsider/2010/05/nsf-to-ask-every-grant-applicant.html (accessed 13 May 2010).

Mesirov JP (2010) Accessible reproducible research. *Science*, **327**, 415.

Michener WK, Brunt JW, Helly JJ, Kirchner TB, Stafford SG (1997) Nongeospatial metadata for the ecological sciences. *Ecological Applications*, **7**, 330–342.

Nelson B (2009) Empty archives. *Nature*, **461**, 160–163.

Overpeck JT, Meehl GA, Bony S, Easterling DR (2011) Climate data challenges in the 21st century. *Science*, **331**, 700–702.

Parmesan C, Yohe G (2003) A globally coherent fingerprint of climate change impacts across natural systems. *Nature*, **421**, 37–42.

Parmesan C, Ryrholm N, Stefanescu C *et al.* (1999) Poleward shifts in geographical ranges of butterfly species associated with regional warming. *Nature*, **399**, 579–583.

Parr CS, Cummings MP (2005) Data sharing in ecology and evolution. *Trends in Ecology and Evolution*, **20**, 362–363.

Piwowar HA, Day RS, Fridsma DB (2007) Sharing detailed research data is associated with increased citation rate. *PLoS ONE*, **2**, e308.

Piwowar HA, Becich MJ, Bilofsky H, Crowley RS; on behalf of the caBIG Data Sharing, Workspace IC (2008) Towards a data sharing culture: recommendations for leadership from academic health centers. *PLoS Medicine*, **5**, e183.

Piwowar HA, Vision TJ, Whitlock MC (2011) Data archiving is a good investment. *Nature*, **473**, 285.

Raff H (2003) A suggestion for the multiple author issue. *Science*, **302**, 55–57.

Raymond ES (1999) *The Cathedral and the Bazaar*. O'Reilly & Associates, Inc., Sebastopol, CA.

Reaman GH (2002) Pediatric oncology: current views and outcomes. *Pediatric Clinics of North America*, **49**, 1305.

Reichman OJ, Schildhauer MP, Jones MB (2011) Challenges and opportunities of open data in ecology. *Science*, **331**, 703–705.

Rezende EL, Lavabre JE, Guimaraes PR, Jordano P, Bascompte J (2007) Non-random coextinctions in phylogenetically structured mutualistic networks. *Nature*, **448**, 925–928.

Roberts L (2002) A tussle over the rules for DNA data sharing. *Science*, **298**, 1312–1313.

Root TL, Price JT, Hall KR, Schneider SH, Rosenzweig C, Pounds JA (2003) Fingerprints of global warming on wild animals and plants. *Nature*, **421**, 57–60.

Root TL, MacMynowski DP, Mastrandrea MD, Schneider SH (2005) Human-modified temperatures induce species changes: joint attribution. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 7465–7469.

Rosenzweig C, Karoly D, Vicarelli M *et al.* (2008) Attributing physical and biological impacts to anthropogenic climate change. *Nature*, **453**, 353–357.

Savage CJ, Vickers AJ (2009) Empirical study of data sharing by authors publishing in PLoS journals. *PLoS ONE*, **4**, e7078.

Schofield PN, Bubela T, Weaver T *et al.* (2009) Post-publication sharing of data and tools. *Nature*, **461**, 171–173.

Sih A, Bell AM, Kerby JL (2004) Two stressors are far deadlier than one. *Trends in Ecology and Evolution*, **19**, 274–276.

Sir Alister Hardy Foundation for Ocean Science (2012) CPR Bibliography. Available at: http://www.sahfos.ac.uk/research/publications/cpr-bibliography.aspx (accessed 12 February 2012).

Smith VS (2009) Data publication: towards a database of everything. *BMC Research Notes*, **2**, 113.

Stein LD (2008) Wiki features and commenting – towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nature Reviews Genetics*, **9**, 678–688.

Stokstad E (2011) Network science: open-source ecology takes root across the world. *Science*, **334**, 308–309.

Strasser CRC, Michener W, Budden A, Koskela R (2011) Dataone: promoting data stewardship through best practices.

Tenopir C, Allard S, Douglass K *et al.* (2011) Data sharing by scientists: practices and perceptions. *PLoS ONE*, **6**, e21101.

Unguru Y (2011) The successful integration of research and care: how pediatric oncology became the subspecialty in which research defines the standard of care. *Pediatric Blood and Cancer*, **56**, 1019–1025.

Wall DH, Bradford MA, St John MG *et al.* (2008) Global decomposition experiment shows soil animal impacts on decomposition are climate-dependent. *Global Change Biology*, **14**, 2661–2677.

Walther GR, Beissner S, Burga CA (2005) Trends in the upward shift of alpine plants. *Journal of Vegetation Science*, **16**, 541–548.

Weltzin JF, Belote RT, Williams LT, Keller JK, Engel EC (2006) Authorship in ecology: attribution, accountability, and responsibility. *Frontiers in Ecology and the Environment*, **4**, 435–441.

Whitlock MC (2011) Data archiving in ecology and evolution: best practices. *Trends in Ecology and Evolution*, **26**, 61–65.

Whitlock MC, McPeek MA, Rausher MD, Rieseberg L, Moore AJ (2010) Data archiving. *The American Naturalist*, **175**, E45–E146.

Zimmerman AS (2008) New knowledge from old data – the role of standards in the sharing and reuse of ecological data. *Science Technology and Human Values*, **33**, 631–652.