esa

ECOSPHERE

The fractured lab notebook: undergraduates and ecological data management training in the United States

C. A. STRASSER[1],† AND S. E. HAMPTON

National Center for Ecological Analysis and Synthesis, University of California Santa Barbara, 735 State Street, Santa Barbara California 93101 USA

Citation: Strasser, C. A., and S. E. Hampton. 2012. The fractured lab notebook: undergraduates and ecological data management training in the United States. Ecosphere 3(12):116. http://dx.doi.org/10.1890/ES12-00139.1

Abstract. Data management is a timely and increasingly important topic for ecologists. Recent funder mandates requiring data management plans, combined with the data deluge that faces scientists, make education about data management critical for any future ecologist. In this study, we surveyed instructors of general ecology courses at 48 major institutions in the United States. We chose instructors at institutions that are likely to train future ecologists, and therefore, are most likely to influence the trajectory of data management education in this field. The survey queried instructors about institution and course characteristics, the extent to which data-related topics are included in their courses, the barriers to their teaching these topics, and their own personal beliefs and values associated with data management and stewardship. We found that, in general, data management topics are not being covered in undergraduate ecology courses for a wide range of reasons. Most often, instructors cited a lack of time and a lack of resources as barriers to teaching data management. Although data are used for instruction at some point in the majority of the courses surveyed, good data management practices and a thorough understanding of the importance of data stewardship are not being taught. We offer potential explanations for this and suggestions for improvement.

Key words: data; data management; ecology; education; environmental sciences; undergraduate.

Received 11 May 2012; revised 15 October 2012; accepted 18 October 2012; final version received 28 November 2012; **published** 21 December 2012. Corresponding Editor: C. D'Avanzo.

[1] Present address: California Digital Library, University of California Office of the President, 415 20th Street, Oakland California 94612 USA.

† **E-mail:** carly.strasser@ucop.edu

INTRODUCTION

Data: a timely topic for ecology

The importance of detailed, accurate record keeping has long been championed in all fields of science, and the traditional physical embodiment of this value on detail is the laboratory notebook. Within the famous notebooks of historical figures such as Darwin, Galileo, and Da Vinci, the reader finds both data and theory, interwoven and exhaustively described. These notebooks, combined with correspondence and publications, created a record of the development of a scientific idea and the foundation of scientific progress (e.g., Costa 2009).

We might expect that scientific records are growing richer as science becomes digitized; however, scientific data and the descriptions of those data (metadata) are often decoupled in the digital age. Data management is more difficult, not easier, when spreadsheets are used to manage data, while metadata are maintained elsewhere—in hand-written notebooks, electronic documents, scraps of paper, etc. In essence, the

traditional laboratory notebook has been fractured, with data and metadata spread across multiple disconnected electronic formats and hardcopies (Butler 2005).

The need for modernization of data management practices in ecology and other disciplines has received increasing attention. An array of tools for managing data and metadata have emerged to meet this need; examples include online laboratory notebook systems and scientific workflow systems (Barseghian et al. 2010, Jones and Gries 2010) and software tools for creating metadata such as Morpho for Ecological Metadata language (Jones et al. 2001). An upswing in concerns about data, its management, and how to properly share and archive it, may be attributed to the National Science Foundation's new requirement that a two-page data management plan be submitted as a supplement to each proposal. Scientists are now required to describe the data they will collect, the policies they will adhere to, and how they will manage and archive it. Of course, the NSF requirement may also be a response to the fact that managing and archiving digital data is an increasingly important task for scientists, and that scientists should think about their strategies early in a project.

There are other motivators for scientists to educate themselves about data management. Journals and publishers are now exploring requirements for data sharing as a condition of publication (Ellison 2010, Whitlock 2010). Big data and the data deluge are flooding scientists with more data than they can process (Maurer et al. 2000, Carlson 2006; Hampton et al., *in press*), and certainly more than they can print out and staple into their lab notebooks. Further, there are calls for more openness in science in general, especially in light of recent scientific scandals related to reproducibility (Brumfiel 2010, Lancet Editorial 2010, Nature Editorial 2010, Pennisi 2010).

Given these motivators, scientists are seeking assistance in properly managing, storing, and archiving their data. We were interested in how this is translating into education of future ecologists. As we transition to an era of better digital data management, are up-and-coming scientists being trained in best practices? Are they learning about data, metadata, and reproducibility? In this study, we sought answers to these questions by surveying instructors of undergraduate ecology courses at institutions likely to be training future ecology graduate students.

Focus on ecology

We chose to focus on the ecology discipline in this study for several reasons. First, ecologists have been known to resist changes to their traditional methods and training (Aronova et al. 2010). Second, there is a small culture of data sharing and archiving in ecology, compared to disciplines such as genomics, physics, and other sciences (McCain 1991, Nelson 2009, Hampton et al. 2012; Hampton et al., *in press*). Ecological data are diverse, consisting of many small, unique data sets that were collected using varied methods (NRC 1995, Bowker 2000, Michener et al. 2007, Zimmerman 2007). This lack of standardization makes data harder to interpret and integrate (Zimmerman 2008) and more costly to manage, but should not be considered sufficient reason to avoid data sharing. Ecology is increasingly participating in the digital information age, and this trend will only continue.

From the supply side, there is increasing availability of online data. The Long Term Ecological Research (LTER) Network houses over 6,000 datasets—a resource that was not available a decade ago when many of today's instructors were undergoing their training (Peters 2010, Porter 2010, Michener et al. 2011). Projects such as the National Ecological Observatory Network (NEON) and the U.S. Integrated Ocean Observing Systems (Baptista et al. 2008) are evidence of growing interest in large-scale data that can only be properly managed using cloud computing and databases and will require sophisticated computing skills. In order to participate in the future of the discipline, ecologists must be capable of documenting their data in standardized formats, creating machine-readable metadata that conform to their discipline's standards, and making their data publicly accessible (Hampton et al. 2012). Digital data manipulation, analysis, and management have become a required basic research skill for all ecologists, similar to writing a coherent sentence.

Current state of data management education

There is relatively little access to training on how to produce and document data sets so that

others can find, understand, and re-use them (Cook et al. 2001). Better education on these topics will go a long way towards instilling in future scientists a deeper appreciation for the value of information about data (metadata) (Michener 2006). The scientific notebook is generally a part of undergraduate education, as are the scientific method and concepts such as reproducibility, but few courses exist that are exclusively devoted to data management practices for scientists. Spreadsheets are often used as the basis of data collection and education; but this is potentially problematic since spreadsheets typically do not promote good data management practices (Jones et al. 2006). The features of spreadsheets that make them desirable for the average researcher, such as extensibility, use of formatting for organization, embedding charts, make them undesirable for preparing data for long-term archiving and re-use. Despite these drawbacks, spreadsheets are the most commonly used software tool in undergraduate ecology programs.

Although digital data has existed for decades, management of those datasets is only now becoming a matter of discussion among researchers. Also lagging is a plethora of educational materials related to data management. There are, however, some resources available for instructors interested in incorporating data management in their curricula. These resources are being generated primarily by institutional libraries (e.g., UC Berkeley Libraries 2011), discipline-specific organizations (e.g., education modules from The Federation of Earth Science Information Partners, 2012; wiki.esipfed.org), and large funded initiatives (e.g., education modules from DataONE, 2012; http://www.dataone.org). There are also data-based exercises being created by organizations such as the Ecological Society of America in their publication *Teaching Issues and Experiments in Ecology* (2012; tiee.esa.org), which integrates overarching ecological and scientific concepts with real datasets and their analysis.

An editorial in *Nature* (2009) summarized the problem best:

> "Universities and individual disciplines need to undertake a vigorous programme of education and outreach about data. Consider, for example, that most university science students get a reasonably good grounding in statistics. But their studies rarely include anything about information management—a discipline that encompasses the entire life cycle of data, from how they are acquired and stored to how they are organized, retrieved and maintained over time. That needs to change: data management should be woven into every course in science, as one of the foundations of knowledge."

There is evidence in the education literature that university-level students would benefit from thinking about data management and organization earlier in their science education rather than later. In fact, both the Benchmarks for Science Literacy (American Association for the Advancement of Science 1993) and the National Science Education Standards (National Committee on Science Education Standards and Assessment, National Research Council 1996) cite data collection, organization, and analysis as crucial pieces in the education of K–12 students; it logically follows that university-level students in their first or second year should be instructed on the next steps related to data collection, which is proper handling of those data. Leonard (2002) stated that "Being able to make accurate observations, predictions, collect and organize data and make inferences are among the most basic of such skills ... for the average citizen who is trying to participate or to survive in a technological society". It is therefore not sufficient to merely understand how to collect data; part of understanding how science works is being able to also organize (i.e., manage) those data.

To this end, we used an in-depth survey of ecology instructors at US institutions likely to be training future ecologists. We queried instructors about their institutions, the ecology course they teach, their beliefs about the importance of data management education, and their personal practices related to data stewardship. Overall, we found data management education to be deficient in the institutions we surveyed. Barriers identified by instructors were primarily associated with a lack of time, but instructors also cited lack of resources, lack of knowledge, and their belief that data management is not an appropriate topic for the level of students they teach.

Methods

Institution selection

To examine course instruction on data management in ecology courses, we selected instruc-

tors from universities and colleges most likely to be teaching future graduate students in ecology. The list of prospective schools was created using three main methods. First, we used the US News and World Report's "Best Graduate Schools" website to generate a list of the top ten Ecology and Evolutionary Biology graduate schools in the US in 2010 (US News and World Report website, 2011). Second, we used the website www.PhDs.org to collate a list of graduate schools in ecology with high National Research Council (NRC) Quality Measures (National Research Council 2009). To do this, we set the NRC Quality Measure at priority five out of a possible five (highest possible priority), with all other priorities set to zero (not considered). Third, we used the same method as above, except the priority was set to five for Research Quality of the institution. Fourth, we obtained a list of the NSF Graduate Research Fellowship recipients for 2010 to 2006 for the life sciences. We removed the following areas of study: biochemistry, biophysics, cell biology, computational biology, developmental biology, genetics, immunology, molecular biology, neurosciences, and nutrition. This left 806 awards. We tallied the number of awards per institution, then used the Carnegie Classification system to determine whether schools were Research Universities (RU) or Baccalaureate/Arts and Sciences (BAS) institutions. All BAS institutions with more than four awards were used in the survey. Based on the four methods above, we generated a list of 51 target institutions for our survey (Appendix A).

Course and instructor selection

We extensively searched each institution's website to determine which course fit our survey best. Fit was determined subjectively based on the course description, course requirements for ecology-related majors, and the academic department(s) housing the course. We focused on undergraduate courses that covered basic topics in Ecology; this was determined by examining syllabi, and in some cases, course materials. Courses required for Ecology/Evolution/Biology majors were given careful scrutiny. Often the course name was indicative of its fit: "General Ecology", "Introduction to Ecology and Evolution", and "Principles of Ecology" were often names of the most appropriate courses.

After the most appropriate course was identified subjectively, we contacted the department to verify that the course we selected was appropriate for our survey. The person contacted was the department chair, the undergraduate course advisor, or a professor within the department. In the email sent, we described the study, identified the potentially relevant course, and asked the contact person to verify that this course was the most appropriate for the survey. We also asked who should be contacted about course content, which in all cases corresponded to the primary instructor. The email sent to the departmental contact person is in Appendix B. Based on Internet research and correspondence with departmental personnel, a list of schools, courses, and contact persons for each course was compiled.

Survey design and implementation

The overarching goal of the survey was to understand the extent to which undergraduates are learning about data management; we were interested in what factors affect this, including course, institution, and instructor characteristics. To accomplish this, the survey consisted of 33 questions that fell into four categories: (1) basic course characteristics, such as class size, laboratory components, reading materials, and prerequisites; (2) the extent to which data management is covered and use of data in the course; (3) instructor opinion about the importance of data management education for undergraduates and perceived barriers to teaching topics related to data management; and (4) instructor characteristics, including year of PhD, percentage of time teaching versus conducting research, and data sharing practices. The full survey may be found in Appendix F.

The survey was conducted online using Survey Monkey (www.surveymonkey.com). An email was sent to the instructors previously identified, with an introduction to the study and a unique link to the survey so it was possible to track which instructors had completed the survey. Emails were sent out 29 March 2011 (Appendix C); the survey was closed 25 May 2011.

Survey processing

We sent emails to 63 instructors at 51 different institutions; we contacted multiple instructors for

Table 1. BAS and RU institution characteristics, course size, and instructors. All values are reported as % except "Years since receiving PhD". For values reported as % (n), n is the number of instructors reporting "yes".

Category	Answer	BAS	RU	Both
Institution and course characteristics	Percent of higher level students (average)	39	63	58
	The course is required	40 (4)	79 (30)	71 (34)
	The lab is mandatory/incorporated into course	90 (9)	42 (16)	52 (25)
	Other courses cover data management	100 (10)	71 (27)	77 (37)
Course enrollment	<50 students	50 (5)	29 (11)	33 (16)
	50–100 students	30 (3)	13 (5)	17 (8)
	>100 students	20 (2)	58 (22)	50 (24)
Instructor characteristics	Years since receiving PhD (average)	20.8	18.3	19.7
	Published in last two years	90 (9)	94 (34)	93 (43)
	Encouraged to share data	100 (9)	100 (34)	100 (43)
	Share data	67 (6)	88 (30)	84 (36)
	Reuse data	44 (4)	79 (26)	71 (30)
Instructor time allotment	Teaching	63	37	45
	Research	25	44	38
	Service	13	19	17

the same course at some institutions to maximize our chances of obtaining at least one response from each institution. We obtained survey responses from 54 instructors at 48 institutions. Our survey unit was the ecology course, so we randomly eliminated duplicates for the same institution, which means that in some cases returned surveys were not used. Of the 48 institutions, 38 were classified as RU institutions and 10 were classified as BAS institutions. To maintain some degree of anonymity for the respondents, we do not report which of the institutions from Appendix A were included in the final survey results.

We downloaded all survey responses and used an R script to clean the data. This process involved removing data from duplicate institutions, deleting data columns not used in our analyses, adding a column for school type (RU versus BAS), and replacing text answers with a numeric coding system. We also trimmed the dataset to remove unused columns and combine some of the columns of data. Standard statistical summaries were then performed on the survey data. The cleaned and trimmed data set and code for cleaning and trimming are available as a Supplement.

Results

Instructor and course characteristics

We asked the course instructors surveyed a series of questions about their courses, position, research, and data practices to determine whether these factors affected data management education in ecology courses (Table 1). The average number of years since a PhD was received was 20, and 93% of the respondents were active researchers (determined by whether they had published a manuscript in a peer-reviewed journal in the last two years).

Of those instructors who were active researchers, 100% stated that they had been encouraged by institutions, journals, or funders to share data, and 84% stated that they had shared data at some point in their careers. This includes peer-to-peer sharing, which is not considered publicly available data. Of the active researchers, 71% reported that they have reused data from others at some point.

We asked active researchers to indicate how much they valued data management and related skills as researchers (Appendix D: Table D1; Fig. 1). In general, active researchers highly valued good data management practices, reuse and sharing of data, and reproducibility.

The amount of time allotted for teaching versus research differed between RU and BAS institutions; on average instructors at RU institutions spend more time teaching and performing "service" duties, and less time on research (Table 1).

Data and data management education in the course

We asked instructors about 12 topics related to data management (Table 2); we determined whether the topic was formally covered in the curriculum, a part of the formal course assess-

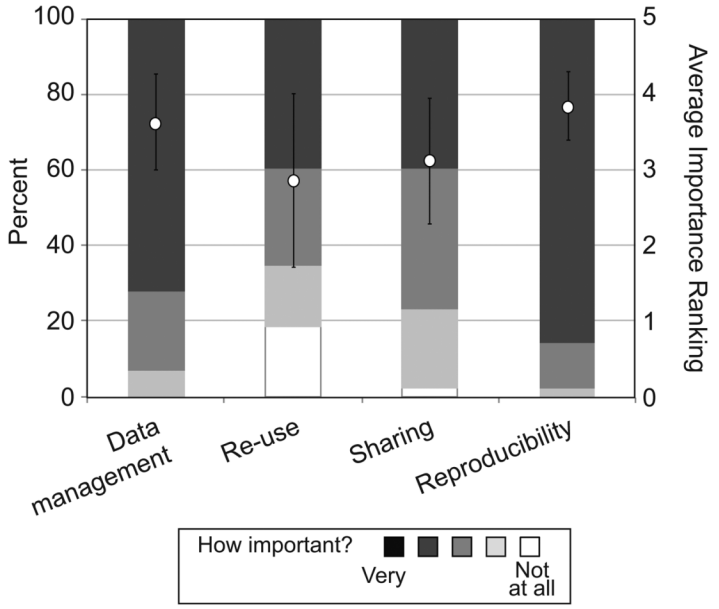

Fig. 1. Instructor rankings of how important each of the data management topics or techniques is to their research. Shaded bars report what percentage of instructors polled answered with that level of importance; darker shades indicate more importance. White data points are the average importance reported by instructors, on a scale of one to five, ±SE.

ment, or both. Coverage of data management topics was idiosyncratic (Fig. 2). Issues of quality assurance were the most commonly taught, appearing in 42% of courses in some form.

In addition to determining whether the 12 topics in Table 2 were covered in the ecology course, we also surveyed instructors about other data-related topics and materials in the course. While most of the courses use student-generated data, few (23%) of them require that students keep notebooks, and the notebooks are evaluated in less than half of those cases (Fig. 3).

Instructors were asked whether the 12 data-related topics were covered in other courses, and therefore not being overlooked in the overall curriculum for ecology students. Most of the instructors listed courses available at their institution that are likely to address some of these topics; these other courses are grouped and reported in Table 3. The most common courses

Table 2. Data management topics.

Topic	Description
Quality control and quality assurance	Making sure that data are accurate and there are no missing values or errors
Naming computer files	Assigning descriptive file names that indicate spatial and/or temporal information about the data
Types of files and software to use	
Metadata generation	Descriptive information describing data content, context, quality, structure, etc.
Workflows	Detailed description, flow chart, or computer script of how raw data were transformed into final results
Protecting data	Backing up data, creating multiple copies in multiple locations
Databases and data archiving	
Data re-use	Using data that was collected for one purpose, for a new or different purpose
Meta-analysis	Statistical synthesis of results of separate studies
Data sharing	Making data (raw or processed) available for use by others
Reproducibility	Making sure results are reproducible, which requires data, metadata, and analytical information
Notebook protocols (lab or field)	

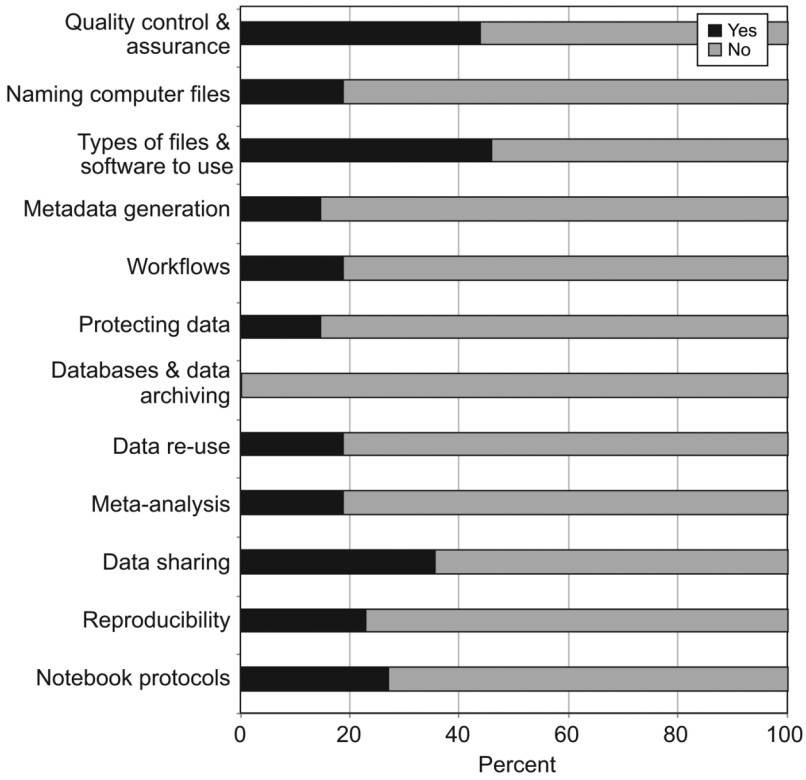

Fig. 2. Percent of ecology courses that address and/or teach the data management topics listed. Original data in Appendix D: Table D2.

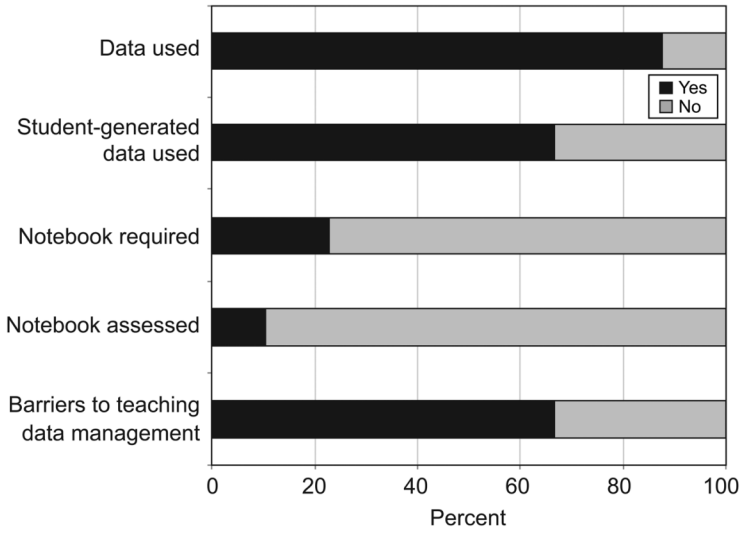

Fig. 3. Percent of ecology course instructors that answered "yes" when asked about the inclusion of each of the data topics in their courses, and whether they perceived barriers to teaching the 12 data management topics in Table 2. Appendix D: Table D2 contains the original data.

Table 3. Other courses offered that cover data management topics.

Course	Frequency mentioned
Ecological Methods and Field Ecology	8
Advanced Ecology	7
Ecology Laboratory	7
Statistics	7
Microbial/Invertebrate/Disease Ecology	6
Aquatic/Marine Biology	6
Plant Ecology	5
Population/Community Ecology	4
Introduction to Biology	4
Theoretical Ecology	2

mentioned were ecology laboratories, advanced ecology courses, or statistics courses.

Professor perceptions

Instructors were asked to rate how important each of the 12 data management education topics were for their students (Appendix D: Table D3; Fig. 4): "1" was designated "not important at all" and "5" was designated "very important". On average, instructors thought data management topics were moderately important: all averages were between 2.3 and 3.3.

We were interested in whether the value an instructor placed on data management in their own research influenced their opinions about the importance of data management topics for undergraduates. To answer this, we calculated the average level of importance each instructor reported for the four data management topics (Appendix D: Table D1) and plotted these values against the average reported importance for undergrads of the 12 data management topics (Appendix D: Table D3). We found a significant weak correlation ($r = 0.306$, $p < 0.05$) between these two factors. In general, instructors who valued data management in their own research also found it important for undergraduates (Fig. 5).

Barriers to data management education

Instructors were asked whether they perceived any barriers to covering the eight data management topics in Table 2. Of the instructors surveyed, 71% perceived barriers to teaching at least one of the eight. We asked those who answered "yes" to elaborate on those barriers in an open-ended response. We received 37 descriptions of perceived barriers and identified eight common barriers that emerged from respondents

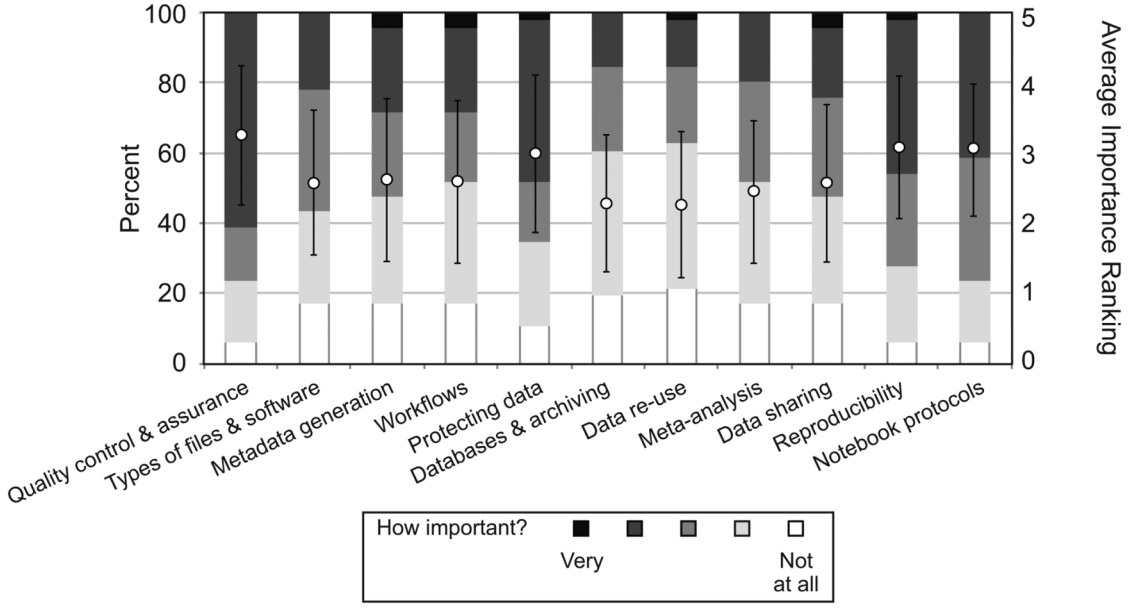

Fig. 4. Instructor rankings of how important each of the data management topics or techniques is for undergraduates in their ecology course. Shaded bars report what percentage of instructors polled answered with that level of importance; darker shades indicate more importance. White data points are the average importance reported by instructors, on a scale of one to five, ±SE.

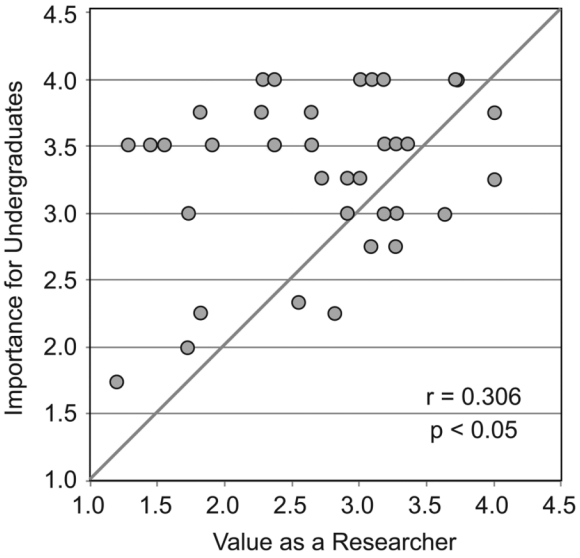

Fig. 5. Comparison of instructors' perceptions of the importance of data management topics for undergraduate students versus how much those instructors value data management in the course of their research; we found a significant weak correlation ($r = 0.306$, $p < 0.05$). The line is the 1:1 line, indicating where importance and value were rated equally.

(Appendix E; Fig. 6). In order of frequency, these barriers included (1) limited time; (2) the topics were not appropriate at the course's level; (3) the topics were or should be covered in a lab section; (4) students in the course did not have the necessary quantitative or statistical skills to cover the topics; (5) lack of funding or resources; (6) the course was too large to cover these topics well; (7) the instructor was not knowledgeable in these topics; and (8) the topics were/should be covered in other courses.

Prediction of data management coverage

We used multiple regression to determine if we could predict whether data management is incorporated into an ecology course based on (1) the instructor's sense of its importance for students and (2) the instructor's perception of barriers to teaching data management topics. To do this, we first summed up how many of the topics in Table 2 were incorporated into each course (Fig. 2). We also summed the importance score across topics provided by each course's instructor (Fig. 4, Table 3).

The importance that researchers ascribed to

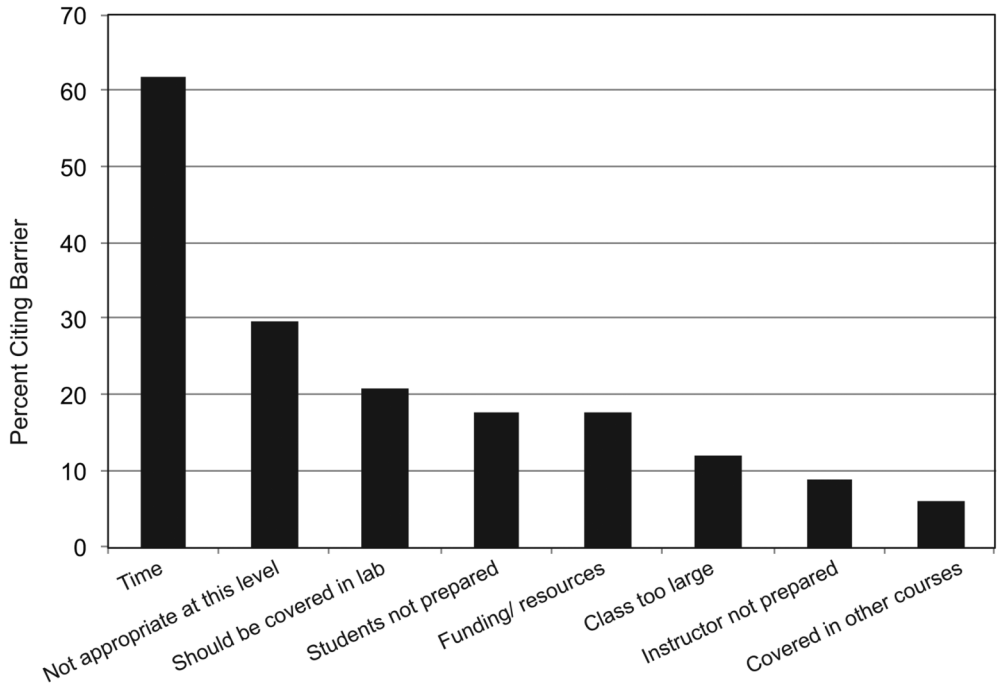

Fig. 6. Percent of instructors who cited the issues listed as barriers to teaching data management topics in Table 2 in a free-form response.

data management topics in their own research was weakly but positively related to the amount of data management material they covered in their courses (Full Model $R^2 = 0.28$; $p = 0.003$; Importance $p = 0.003$), with no correlations evident for Barriers ($p = 0.45$) or the interaction ($p = 0.15$).

Discussion

A lack of data education

Based on results from our survey, data management education is not currently a priority for ecology instructors teaching general ecology courses at the undergraduate level. Less than half of the courses surveyed covered any of the 12 data management topics in Table 2. Few instructors of these courses required that students keep laboratory or field notebooks, although student-generated data was used in more than 50% of the ecology courses. These results indicate that undergraduates in ecology are not learning how to properly document, organize, and archive data sets in the context of their relevance to practicing ecology. Notably, 77% of instructors indicated that data management should be taught in a different course; this may imply that although the ecology professors think these topics are important, they are not willing or able to cover them in the courses they teach.

Potential causes

The lack of data management education that we found cannot be easily attributed to a single characteristic of the institution, the course, or the instructor. Although there are complex reasons why data management is being overlooked in curriculum and assessment at the undergraduate level, the two most prominent factors influencing this phenomenon are the instructors' perceptions of the importance of data management for the undergraduates in their courses and these instructors' struggles with limits on time and resources.

Instructors are managing large class sizes with limited personnel and resources, combined with their need to attend to research and service duties. Based on self-reported barriers to teaching, the majority of the instructors surveyed (over 60%) struggled with the amount of time they were given to cover a broad range of ecological topics, theories, and concepts (Fig. 6).

Before conducting our survey, we hypothesized that there would be differences in what was being taught about data management at Baccalaureate/Arts and Sciences institutions (BAS) versus research universities (RU). However, differences did not readily emerge, suggesting that the problem is systemic across institutions.

Conclusions

There are two major conclusions we can draw from the survey. First, there is an urgent need for trained scientists who are instructing future scientists to understand the importance of data management education. Responses to our survey suggest that instructors themselves are not necessarily well educated in the importance of data management and good data stewardship. Instructors recognize the value of good data management, data sharing, and reproducibility in their own research, but are not transferring that knowledge to their undergraduate students. Second, instructors teaching undergraduate ecology courses need to be made aware of materials available to help them incorporate data management topics into their curriculum so that students understand the importance of good data management practices in being a practicing ecologist. Groups such as DataONE, Data Conservancy, the US Geological Survey, and professional societies such as the Ecological Society of America are working to create comprehensive sets of materials for educating undergraduates and graduate students in data management.

Incorporation of data in the classroom through projects with designs similar to the distributed knowledge network (Andelman et al. 2004) and CENS data in the classroom (Wallis et al. 2006) will provide students with a much deeper understanding of data analysis and the scientific process in general. Instructors might also consider having students maintain electronic laboratory notebooks rather than paper notebooks (Butler 2005); this encourages digitization of notes and materials earlier in the course of data collection, and therefore facilitates conversations about backing up and protecting data, and other best practices for data management (Borer et al. 2009). When data analysis exercises and new technologies for data management are integrated in the classroom or laboratory, students will be better prepared to enter an age in which ecologists and

other professionals are expected to work with increasing levels of data sophistication. Not only will ecology be better served, but so will society in general—skills in data management and analysis are highly transferable and of increasing demand across disciplines and sectors.

Acknowledgments

This work was funded by DataONE (NSF Grant No. OCI 0830944). The survey for this study was administered in accordance with the policies of University of California at Santa Barbara's Human Subjects Committee, Office of Research. J. Parker and D. Marsh were instrumental in survey design; B. Grant helped with survey conception. J. Tewksbury, J. Williams, S. Queenborough, G. Mittelbach and K. Gross provided valuable feedback. J. Byrnes assisted with statistical brainstorming. We are extremely grateful for comments on the manuscript provided by C. Duke, C. Tenopir, K. Douglass, and J. Porter.

Literature Cited

American Association for the Advancement of Science. 1993. Benchmarks for science literacy. AAAS Publication, New York, New York, USA.

Andelman, S., C. Bowles, M. Willig, and R. Waide. 2004. Understanding environmental complexity through a distributed knowledge network. BioScience 54:240–246.

Aronova, E., K. S. Baker, and N. Oreskes. 2010. Big science and big data in biology: From the International Geophysical Year through the International Biological Program to the Long Term Ecological Research (LTER) Network, 1957–Present. Historical Studies in the Natural Sciences 40.

Baptista, A., B. Howe, J. Freire, D. Maier, and C. T. Silva. 2008. Scientific exploration in the era of ocean observatories. Computing in Science and Engineering 10:53–58.

Barseghian, D., I. Altintas, M. B. Jones, D. Crawl, N. Potter, J. Gallagher, P. Cornillon, M. Schildhauer, E. T. Borer, E. W. Seabloom, and P. R. Hosseini. 2010. Workflows and extensions to the Kepler scientific workflow system to support environmental sensor data access and analysis. Ecological Informatics 5:42–50.

Borer, E., E. Seabloom, M. Jones, and M. Schildhauer. 2009. Some simple guidelines for effective data management. Bulletin of the Ecological Society of America 90:205–214.

Bowker, G. C. 2000. Biodiversity datadiversity. Social Studies of Science 30:643–683.

Brumfiel, G. 2010. Scientists question cancer gene trials at Duke University. Shots: NPR's Health Blog 20 July 2010. http://www.npr.org/blogs/health

Butler, D. 2005. A new leaf. Nature 436:20–21.

Carlson, S. 2006. Lost in a sea of science data. Chronicle of Higher Education 52:A35. http://chronicle.com/weekly/v52/i42/42a03501.htm

Cook, R., R. Olson, P. Kanciruk, and L. Hook. 2001. Best practices for preparing ecological data sets to share and archive. Bulletin of the Ecological Society of America 82:138–141.

Costa, J. T. 2009. The Darwinian revelation: tracing the origin and evolution of an idea. BioScience 59:886–894.

Ellison, A. M. 2010. Repeatability and transparency in ecological research. Ecology 91:2536–2539.

Hampton, S., C. Strasser, J. Tewksbury, W. Gram, A. Budden, A. Batcheller, C. Duke, and J. Porter. In press. Big data and the future for ecology. Trends in Ecology & Evolution.

Hampton, S. E., J. J. Tewksbury, and C. A. Strasser. 2012. Ecological data in the Information Age. Frontiers in Ecology and the Environment 10:59–59.

Jones, M. B., C. Berkley, J. Bojilova, and M. Schildhauer. 2001. Managing scientific metadata. IEEE Internet Computing 5:59–68.

Jones, M. B., and C. Gries. 2010. Advances in environmental information management. Ecological Informatics 5:1–2.

Jones, M. B., M. P. Schildhauer, O. J. Reichman, and S. Bowers. 2006. The new bioinformatics: Integrating ecological data from the gene to the biosphere. Annual Review of Ecology Evolution and Systematics 37:519–544.

Lancet Editorial. 2010. Retraction: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. Lancet 375:445.

Leonard, W. 2002. How do college students best learn science? Chapter 2 *In* Innovative techniques for large-group instruction. National Science Teachers Association Press, Arlington, Virginia, USA.

Maurer, S. M., R. B. Firestone, and C. R. Scriver. 2000. Science's neglected legacy. Nature 405:117–120.

McCain, K. W. 1991. Communication, competition, and secrecy: the production and dissemination of research-related information in genetics. Science, Technology, & Human Values 16:491–516.

Michener, W. 2006. Meta-information concepts for ecological data management. Ecological Informatics 1:3–7.

Michener, W. K., J. H. Beach, M. B. Jones, B. Ludaescher, D. D. Pennington, R. S. Pereira, A. Rajasekar, and M. Schildhauer. 2007. A knowledge environment for the biodiversity and ecological sciences. Journal of Intelligent Information Systems 29:111–126.

Michener, W. K., J. Porter, M. Servilla, and K.

Vanderbilt. 2011. Long term ecological research and information management. Ecological Informatics 6:13–24.
National Committee on Science Education Standards and Assessment, National Research Council Board on Science Education. 1996. National science education standards. National Academies Press, Washington, D.C., USA.
National Research Council. 1995. Finding the forest in the trees: The challenge of combining diverse environmental data. National Academies Press, Washington, D.C., USA.
National Research Council. 2009. Assessment of research doctoral programs in the United States. National Academies Press, Washington, D.C., USA.
Nature Editorial. 2009. Data's shameful neglect. Nature 461:145–145.
Nature Editorial. 2010. Climate of suspicion. Nature 463:269–269.
Nelson, B. 2009. Data sharing: Empty archives. Nature 461:160–163.
Pennisi, E. 2010. Discoverer asks for time, patience over arsenic bacteria controversy. Science 330:1734–1735.
Peters, D. P. C. 2010. Accessible ecology: synthesis of the long, deep, and broad. Trends in Ecology & Evolution 25:592–601.
Porter, J. 2010. A controlled vocabulary for LTER datasets. LTER Databits. Information Management Newsletter of the LTER Network May.
UC Berkeley Science Libraries. 2011. Preparing data management plans for NSF grant applications. University of California, Berkeley, California, USA.
US News and World Report. 2011. Best graduate schools: top science schools in ecology. http://grad-schools.usnews.rankingsandreviews.com
Wallis, J. C., S. Milojevic, C. L. Borgman, and W. A. Sandoval. 2006. The special case of scientific data sharing with Education. *In* A. Grove, editor. 69th Annual Meeting of the American Society for Information Science and Technology (ASIST) Volume 43.
Whitlock, M. C. 2011. Data archiving in ecology and evolution: best practices. Trends in Ecology & Evolution 26:61–65.
Zimmerman, A. 2007. Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. Integrated Journal on Digital Libraries 7:5–16.
Zimmerman, A. 2008. New knowledge from old data: The role of standards in the sharing and reuse of ecological data. Science, Technology & Human Values 33:631–652.

Supplemental Material

Appendix A

List of Institutions

Amherst College
Bowdoin College
Brown University
Carleton College
Colorado College
Colorado State University
Columbia University
Cornell University
Dartmouth College
Duke University
Emory University
Evergreen State College
Grinnell College
Harvard University
Indiana University (Bloomington)
Lewis and Clark College
Michigan State University
Middlebury College
Oberlin College
Oregon State University
Princeton University
Purdue University
Rice University
San Diego State University
Stanford University
Swarthmore College
University of Arizona
University of California Berkeley
University of California Davis
University of California Irvine
University of California Riverside
University of California Santa Barbara
University of California Santa Cruz
University of Chicago
University of Colorado
University of Florida
University of Georgia
University of Idaho
University of Illinois Urbana-Champaign
University of Kansas
University of Maryland
University of Michigan
University of Minnesota
University of Montana
University of New Mexico

University of Southern California
University of Texas Austin
Utah State University
Washington University in St. Louis
Williams College
Yale University

Appendix B

Inquiry Email

Dear [department contact's name here],

I am a postdoctoral researcher at the National Center for Ecological Analysis and Synthesis (NCEAS) at UC Santa Barbara. I'm interested in undergraduate Ecology education in data management and reuse.

I'm contacting schools with the top ecology graduate programs to survey their undergrad education in data management, and [institution] is on this list. I'm doing some preliminary fact-checking (not part of the official study) and have a couple of questions I'm hoping you can answer. It should only take a minute or two of your time, and I would very much appreciate your prompt response.

I would like to interview the instructor for the primary undergraduate ecology course at [institution]. The course I am interested in would be advised for undergrads planning to pursue graduate studies in ecology and/or Environmental Sciences.

Here are my questions for you:

- What is the general undergraduate ecology course at [institution], best for undergrads planning to pursue graduate studies in ecology? Based on my web investigation, I assume it would be [my guess here].
- Who is the best person(s) to contact about the curriculum and assessment scheme for the course you named in (1)? If there are multiple sections or instructors, is there one main person to contact?

Thanks so much for your help, and if you would like more information about the study I would be happy to oblige. I could also call you with these questions (or you can call me any time at the number below).

Sincerely,
Carly Strasser

Appendix C

Email to Instructors

Dear Dr. [Instructor name here],

Recently I have been in touch with either you or your department chair about the course [Course name here]. I am conducting a systematic survey of undergraduate ecology courses, and I hope you will help me by answering questions about this course. [Institution] was selected to be a part of this survey because of its strong likelihood of training undergraduate students that might go on to graduate programs in ecology. This survey will contribute to knowledge about data management education for the DataONE Project (www.dataone.org). The goals of the survey are: (1) to better understand the content of undergraduate ecology courses and (2) to assess the potential barriers to teaching undergraduates about topics related to data. The results of the survey will help DataONE to develop effective and relevant educational resources for ecology instructors.

Prior participants report that the survey takes around 20 minutes to complete. I am interested in assuring responses from all 50 of the schools I have contacted, therefore a record of your participation will be maintained. I assure you that I will maintain confidentiality of participants and their institution when reporting survey results.

I recognize that your time is valuable, but I am sincerely interested in learning more about the content of your ecology course and the potential barriers to effectively teaching topics related to data. Please take the time to complete this survey before April 30th.

Here is a link to the survey: [Link here]

This link is uniquely tied to your email address. Please do not forward this message. If you have any questions about the survey or this project, feel free to contact me at the email address or phone number below.

Sincerely,
Carly Strasser

If you think you received this email in error, please follow this opt-out link: [Link here]

APPENDIX D

Summary of Survey Data

Table D1. Instructor rankings of how important each of the data management topics or techniques is to their research. Values are % (n), where n is the number of instructors indicating that ranking.

Topic	Institution type	Not at all	Somewhat	Moderately	Very	Extremely
Good data management practices	BAS	0 (0)	11 (1)	33 (3)	55 (5)	0 (0)
	RU	0 (0)	6 (6)	18 (6)	76 (26)	0 (0)
	Both	0 (0)	7 (3)	21 (9)	72 (31)	0 (0)
Reuse of data	BAS	44 (4)	11 (1)	11 (1)	33 (3)	0 (0)
	RU	12 (4)	18 (6)	30 (10)	41 (14)	0 (0)
	Both	19 (8)	16 (7)	26 (11)	40 (17)	0 (0)
Sharing data	BAS	11 (1)	33 (3)	44 (4)	11 (1)	0 (0)
	RU	0 (0)	18 (6)	35 (12)	47 (16)	0 (0)
	Both	2 (1)	21 (9)	37 (16)	40 (17)	0 (0)
Reproducibility of data	BAS	0 (0)	0 (0)	22 (2)	78 (7)	0 (0)
	RU	0 (0)	3 (1)	9 (3)	88 (29)	0 (0)
	Both	0 (0)	2 (1)	12 (5)	86 (36)	0 (0)

Table D2. Percent of instructors reporting that each of the data management topics listed is covered in their course. Values are % (n), where n is the number of instructors indicating that a topic was covered.

Topic	BAS	RU	Both
Quality control and quality assurance	60 (6)	39 (15)	44 (21)
Naming computer files	30 (3)	16 (6)	19 (9)
Types of files and software to use	80 (8)	37 (14)	46 (22)
Metadata generation	30 (3)	11 (4)	15 (7)
Workflows	40 (4)	13 (5)	19 (9)
Protecting data	20 (2)	13 (5)	15 (7)
Databases and data archiving	0	0	0
Data re-use	40 (4)	13 (5)	19 (9)
Meta-analysis	10 (1)	21 (8)	19 (9)
Data sharing	50 (5)	32 (12)	35 (17)
Reproducibility	30 (3)	21 (8)	23 (11)
Notebook protocols (lab or field)	40 (4)	24 (9)	27 (13)
Notebook required	50 (5)	16 (6)	23 (11)
Notebook assessed if required	20 (1)	67 (4)	45 (5)
Data used in course	100 (10)	84 (32)	88 (42)
Student-generated data used in course	90 (9)	61 (23)	67 (32)

Table D3. Instructor rankings of how important each of the data management topics or techniques is to students in their ecology course. Values are % (n), where n is the number of instructors indicating that ranking. There were no significant differences in the responses from instructors at BAS and RU institutions.

Topic	Institution type	Not at all	Somewhat	Moderately	Very	Extremely
Quality control and quality assurance	BAS	0 (0)	1 (0.1)	1 (0.1)	8 (0.8)	0 (0)
	RU	3 (0.08)	7 (0.19)	6 (0.17)	20 (0.56)	0 (0)
	Both	3 (0.07)	8 (0.17)	7 (0.15)	28 (0.61)	0 (0)
Types of files and software to use	BAS	2 (0.2)	0 (0)	7 (0.7)	1 (0.1)	0 (0)
	RU	6 (.17)	12 (0.33)	9 (0.25)	9 (0.25)	0 (0)
	Both	8 (0.17)	12 (0.26)	16 (0.35)	10 (0.22)	0 (0)
Metadata generation	BAS	2 (0.2)	2 (0.2)	3 (0.3)	2 (0.2)	1 (0.1)
	RU	6 (.17)	12 (0.33)	8 (0.22)	9 (0.25)	1 (0.03)
	Both	8 (0.17)	14 (0.30)	11 (0.24)	11 (0.24)	2 (0.04)
Workflows	BAS	2 (0.2)	5 (0.5)	1 (0.1)	1 (0.1)	1 (0.1)
	RU	6 (.17)	11 (0.31)	8 (0.22)	10 (0.28)	1 (0.03)
	Both	8 (0.17)	16 (0.35)	9 (0.20)	9 (0.20)	2 (0.04)
Protecting data	BAS	1 (0.1)	2 (0.2)	1 (0.1)	6 (0.6)	0 (0)
	RU	4 (0.11)	9 (0.25)	7 (0.20)	15 (0.42)	1 (0.03)
	Both	5 (0.11)	11 (0.24)	8 (0.18)	21 (0.46)	1 (0.21)
Databases and data archiving	BAS	1 (0.1)	6 (0.6)	1 (0.1)	2 (0.2)	0 (0)
	RU	8 (0.22)	13 (0.36)	10 (0.28)	5 (0.14)	0 (0)
	Both	9 (0.20)	19 (0.41)	11 (0.24)	7 (0.15)	0 (0)
Data re-use	BAS	2 (0.2)	5 (0.5)	2 (0.2)	1 (0.1)	0 (0)
	RU	8 (0.22)	14 (0.39)	8 (0.22)	5 (0.14)	1 (0.03)
	Both	10 (0.22)	19 (0.41)	10 (0.22)	6 (0.13)	1 (0.21)
Meta-analysis	BAS	3 (0.3)	6 (0.6)	0 (0)	1 (0.1)	0 (0)
	RU	5 (0.14)	10 (0.28)	13 (0.36)	8 (0.22)	0 (0)
	Both	8 (0.18)	16 (0.35)	13 (0.28)	9 (0.20)	0 (0)
Data sharing	BAS	0 (0)	3 (0.3)	3 (0.3)	3 (0.3)	1 (0.1)
	RU	8 (0.22)	11 (0.31)	10 (0.28)	6 (0.17)	1 (0.03)
	Both	8 (0.18)	14 (0.30)	13 (0.28)	9 (0.20)	2 (0.04)
Reproducibility	BAS	1 (0.1)	1 (0.1)	3 (0.3)	5 (0.5)	0 (0)
	RU	2 (0.06)	9 (0.25)	9 (0.25)	15 (0.42)	1 (0.03)
	Both	3 (0.07)	10 (0.22)	12 (0.26)	20 (0.43)	1 (0.21)
Notebook protocols (lab or field)	BAS	0 (0)	1 (0.1)	5 (0.5)	4 (0.4)	0 (0)
	RU	3 (0.08)	7 (0.19)	11 (0.31)	15 (0.42)	0 (0)
	Both	3 (0.07)	8 (0.17)	16 (0.35)	19 (0.41)	0 (0)

Appendix E

Barriers: Instructor Responses

Below are verbatim responses to the survey question:

"Are there barriers to your teaching any of the following topics? (refer to above question for descriptions if necessary). [See Table 2 for the list of topics]. If you answered "Yes" to any of the above, please describe the barriers you face (e.g., time, not enough information, not appropriate at this level etc.)."

Some text has redacted to maintain privacy; replacement text is found in brackets.

1. I don't have time to teach any of this in lecture; presumably they get some in the lab. But, the lab is not well designed. It isn't really an "ecology" lab, but more of diversity of life.
2. The main barrier is Time. My course is an introductory course and although the topics raised are important, we simply do not have enough time to cover them and I don't think the students are conceptually savvy enough to understand why the topics are important. We cover most of these topics in three advanced course for undergraduates on ecological field research.
3. This is a first undergraduate ecology course. We simply do not have the time or human resources to deal with these issues at this early juncture in the student's training. A few of the students who take on specific independent projects do get exposed to these various issues.
4. time–in one coure [sic] we cover basics of ecological theory and have a heavy field emphasis so that students learn how to be

observant, see patterns, think about how to generate research hypotheses based on these patterns. In small groups, students also write proposals and then complete the work they propose and report the results in a paper and oral presentation. This emphasis on field skills and research skills makes it impoosible [sic] to include many of the skills described in the survey. [I] would be very interested in including some of the topics covered in this survey [in a future course]. If you could provide materials (outlines, drafts, etc.) soon, I will incorporate some of the info into the course as I plan it over the next two months.

5. time (mine and the students')
6. Time is the primary barrier. This course covers population, community, and ecosystem ecology as well as several applications of ecology, and there is simply no time for methods beyond understanding what an experiment is and what replication is and why it is important.
7. Time is the main factor. I would also say that these topics seem to be beyond the level of information this course is designed to cover
8. I am not trained in many of these topics related to management of large datasets and I am learning as I collect the data which is not the best way to figure things out. Quality control is very challenging since it is a manual process prone to mistakes.
9. While I answered "No" to all, time and competing activities are universal constraints. The prime competing activities are: 1. remedial education in natural history, 2. presenting conceptual overviews and testing them with qualitative ecological patterns, 3. teaching experimental/observational design, and 4. learning to gather original data in the field, and 5. actually gathering data in field and lab.
10. The course I teach is an upper division lecture to a large group of students (>150). There is insufficient time to cover most of these topics in the 10-week course.
11. Limited time during course [redacted]. we are on the quarter system and only have ten weeks to cover a wide range of topics. Also, funding cuts means that there is no money to offer a laboratory for the ecology course where such skills can be best addressed.
12. In most ecology classes, there is little time to cover these topics.
13. My class has no lab section - I can't cram it all in/don't have an appropriate venue for it right now My class is large and I have no TAs (thus no lab sections) so some of htis [sic] is difficult to do with 90 students in a lecture hall My class has 50% social science students with no math/science who will take no other math/science so it's not really appropriate for the level/student audience
14. I think adequate treatment of meta-analysis would be in a stats class not an Ecology lab class
15. insufficient course funding for software
16. Insufficient time, lack of a unified approach among colleagues, lack of a course on this topic
17. For all of the topics, there is always a tradeoff among competing priorities.
18. barriers–no associated lab with the course (but there are lab courses that cover this); large lecture class size in auditorium style room; only 10 weeks to cover both ecology and evolution;
19. Time for all marked 'yes'. Seems like a second-level of endeavor, not what you start with in an introductory lab. This is the first biology lab they wlll [sic] have had so we have other first-level objectives.
20. Not appropriate at this level given limited resources (mainly time).
21. Students come to my class with no knowledge of what a mean is, how to calculate it, what uncertainty is. Many students do not know how to make or interpret a graph, much less generate any data. I have been using TIEE (teaching issues and experiments in ecology) modules to help students learn these skills. With 100+ students in a class and no teaching assistants we do not have adequate resources.
22. I am answering these questions as they pertain to our large [*course name*] lecture class, not the optional lab class which I am

also responsible for. Barriers exist in the lecture class to all of these due to its size. I introduce data sets from my own lab that we "analyze" in the lecture hall - this involves me describing the basic principles of regression, for example, to assess whether an apparent trend is meaningful or not. However, how those data are obtained, and how they are processed to generate the tell-tale graph, is not covered, and really can't be if I want to cover what I consider to be critical topics of how human are altering the Earth system.

23. Not appropriate at this level for a non-lab course This is a Principles course providing a broad overview of ecological concepts, not data collection or metadata analysis,. [sic] I would consider all of those topics to be more appropriate for graduate studies.
1) Time; 2) If I was teaching an intensive course I might prioritize some of these lower than learning how to analyze data and present results, so priorities; and 3) I haven't researched it, but it would certainly be useful if there was curriculum available to teach from!
24. time; not appropriate at this level; variation in student familiarity with software; my limited (but developing) awareness of data management and archiving standards in ecology
25. The main barriers are two: First, the course has more than 200 students when I teach it, so individualized assessment and feedback is difficult, and I prioritize other skills (writing, critical thinking) for assessment and feedback. Second, thee is always more that I want to cover than there is time to cover, using a reasonable pace that allows for interaction in the classroom. In addition, much of this is covered in our separate (optional) lab. So in lecture, I focus on the type of data and approach that go into the examples that I present. . . . trying to highlight the pros and cons associated with model simulations, lab experiments, field experiments, field observations, etc.
26. time some of issues overly complex and specific when just introducing students to basics of the field and major concepts
27. Time and appropriateness at this level. The main goal of the class is teaching ecological concepts. I could see a more advanced course covering data usage being a nice addition to our program. However, the students have started taking a longer introductory biology sequence, in which they do a whole semester on ecology and evolutionary biology. So the students are becoming better prepared and it may be possible to integrate these things into General Ecology to a small degree at least.
28. students have limited knowledge of statistics.
29. The math skills of some of our students are not up to speed; there are no math prerequisites for the course
30. course does not have a lab (funding), and the course rotates with another instructor who has a traditional style. It will require significant re-design to include these elements (though worthwhile). Over the longer term I want to include many of these elements.... but also hope to get a lab section.
31. Although I consider that all of the topics that I checked "yes" for are important, in a broad survey course with students at a lot of different levels, I am focusing on much more basic material. I do a fair amount of biostatistics in the ecology lab, but it is more basic. Our students are required to take more statistics and quantitative classes, and I am just stretched so thin that I can't cover all of these. I do feel very strongly, however, that the use of field notebooks is a critical and vanishing skill. Students keep a field notebook for the whole semester, and it is worth a large part of their grade. [text redacted]. Students are given suggestions and examples about how to keep these notebooks before they start. This has been a transformative experience for some students.
32. I answered no to all of these. That is, I could cover these topics—but many are not appropriate for the beginning ecologist. In the Ecology course we want to get our students "into" and excited about the natural world and ecology—so some (but

not all) of these topics would not be appropriate at this level.
33. While I would cover all of these data collection, storing, handling, preserving topics with an undergraduate researcher in my lab, I don't feel they need to be covered in a general ecology course (where most of my students are pre meds) or in a field course where we are focused on learning to collect and analyze primary data. I do talk a lot about database development and meta-analysis [in another course].
34. Time. There isn't enough time to cover the concepts that I would like to cover, let alone QAQC issues, metadata, data archival. Etc. Many of these issues are not very important in a basic ecology course (where many of the students won't go further in actually doing science)—they need to understand the basic issues here, but not the detail.
35. Because of time constraints, we choose not to emphasize the topics that are checked "yes" above.
36. Our emphasis is usually on data analysis and interpretation to emphasize concepts taught in class, and there is not enough time to get into meta-analyses. Analysis and data processing are each a piece of this course, but not the major emphasis; we focus more on learning ecological concepts from the data rather than data care per se. Students generally lack strong quantitative and database skills or interests.
37. No lab for this course

Appendix F

The full survey is available in PDF. http://dx.doi.org/10.1890/ES12-00139.2

Supplement

Raw survey results and data cleaning and processing scripts for R (*Ecological Archives* C003-013-S1).