# Data Information Literacy Case Study Directory

Volume 2 *Cornell University*
Number 1 *Natural Resources*

Article 1

2015

# Natural Resources / Graduate Students / Wright & Andrews / Cornell University / 2013

Sarah J. Wright
*Cornell University*, sjw256@cornell.edu

Camille Andrews
*Cornell University*, ca92@cornell.edu



## Recommended Citation

Wright, Sarah J. and Andrews, Camille (2015) "Natural Resources / Graduate Students / Wright & Andrews / Cornell University / 2013," *Data Information Literacy Case Study Directory*: Vol. 2: No. 1, Article 1.
http://dx.doi.org/10.5703/1288284315476

**DEVELOPING A FOR- CREDIT COURSE TO TEACH DATA INFORMATION LITERACY SKILLS: A Case Study in Natural Resources**

Sarah J. Wright, Cornell University

Camille Andrews, Cornell University

INTRODUCTION

The Cornell University Data Information Literacy (DIL) project team worked with a faculty member and graduate students in natural resources. The faculty member's lab collects data on longitudinal changes in fish species and zooplankton—namely their occurrence, population abundance, growth, and diet—in Lake Ontario. After interviewing the faculty member, a former student, and a lab technician, we determined that the DIL needs for this area were primarily data management and organization and data quality and documentation, including metadata and data description. We also placed a secondary focus on databases and data formats, data visualization and representation, and cultures of practice, including data sharing.

To address these needs, the Cornell team focused on two separate educational tracks. The first was a series of DIL workshops, open to the whole Cornell community, which was an introduction to data management and data management plans (DMPs), data organization, and data documentation. The second was a 6-week credit course on data management for graduate students in natural re- sources taught by the faculty member and the data librarian, Sarah J. Wright, in the spring of 2013. The course built on the previous workshop topics and also included sections on data quality, data sharing, data analysis, and visualization.

Assessment for the workshops involved using post-instruction surveys. The for-credit course assessment included formative "1-minute papers," very short, anonymous exercises per- formed at the end of each class; instructor feed- back on active learning exercises (including an optional DMP exercise graded by a rubric—see Appendix A to this chapter); and a final survey that asked students to self-report on perceived skills before and after taking the class. The feedback was generally very positive, with the majority of students in the credit course indicating that they would recommend it to other graduate students in natural resources. They also reported an increase in their skill levels for all outcomes.

This chapter will discuss the Cornell case study and our instructional approaches. The strengths of our program were that we

- introduced graduate students to major concepts in data management;
- built and gathered modules, exercises, and tools that can be used in a range of educational situations;

• exposed current gaps in data management training;

• allowed students to network and ex- change information;

• built awareness and relationships with faculty.

Ways in which we can improve are to:

• provide more hands-on exercises so that students can apply the skills they learn to their research data;

• tailor the outcomes of the workshops and the course to specific skill levels and other disciplines;

• build and gather more curriculum re- sources and activities at both low- and high-skill levels;

• increase outcomes-based assessment and experiment with ways to make sessions more student-centered and peer-led.

## LITERATURE REVIEW AND ENVIRONMENTAL SCAN OF DATA MANAGEMENT IN NATURAL RESOURCES AND ECOLOGY

The faculty member who worked with our Cornell team has a lab that collects data on longitudinal changes in fish species and zooplankton. This faculty member has long been an advocate of improving the data management skills of graduate students, and therefore was a natural partner for this project. Our faculty member's concern with data management reflected general trends in the larger field of ecology, which has increasingly emphasized data management and curation at both a macro and a micro level. For example, Wolkovich, Regetz, and O'Connor (2012) note:

> Because an ecological dataset [is] collected at a certain place and time [it] represents an irreproducible set of observations. Ecologists doing local, independent research possess . . . a wealth of information about the natural world and how it is changing. Although large-scale initiatives will increasingly enable and reward open science, change demands action and personal commitment by individuals—from students and PIs [principal investigators]. (p. 2102)

A great deal of the literature focused on higher level issues, such as big data, cyberinfrastructure, and the development of metadata standards, or on an individual project as a microcosm of these issues. Given the heterogeneous and interdisciplinary nature of ecological data and the need for integrative studies in areas such as climate change, several authors (Carr et al., 2011; Jones, Schildhauer, Reichman, & Bowers, 2006; Michener & Jones, 2012; Wolkovich et al., 2012) addressed bioinformatics, ecoinformatics, and data sharing writ large, including the current state of the art and the need for better data management and coordination between various areas of ecological research. Others (Gil, Hutchison, Frame, & Palanisamy, 2010; Michener, Brunt, Helly, Kirchner, & Stafford, 1997) explored the variety of metadata standards for ecological data, the need for structured metadata and crosswalks to facilitate integration and interoperability of heterogeneous data sets, and the existing and needed partnership efforts necessary to advance this. In other cases, the literature outlines cyber- infrastructure needs for long-term ecological re- search, including particular technical solutions and issues with data collection, modeling, and management, such as the difficulties of collecting and harvesting heterogeneous data from a network of sites, building cross-searchable digital repositories, and accurately modeling with existing data (Barros, Laender, Gonçalves, Cota, & Barbosa, 2007; Magnusson & Hilborn, 2007; McKiernan, 2004). Institutions such as "The Long Term Ecological Research Network" (2012; Michener, Porter, Servilla, & Vanderbilt, 2011), DataONE (n.d.a), the Knowledge Net- work for Biocomplexity (2005), and for limnology the Global Lake Ecological Observatory Network (n.d.) championed high-level efforts toward providing researchers with centralized repositories, resources, tools, and training to address data management needs. For example, the Ecological Metadata Language (EML) and data management tools such as Morpho from the Knowledge Network for Biocomplexity are standards and tools that are widely available (Fegraus, Andelman, Jones, & Schildhauer, 2005; Knowledge Network for Biocomplexity, n.d.).

Among the natural resources graduate students we interviewed, there was a lack of aware- ness of existing practices, tools, or standard best practices in other areas, as well as a demand for point-of-need information and instruction at a very basic level. Although compilations of basic guidelines exist, such as those published in the Bulletin of the Ecological Society of America (Borer, Seabloom, Jones, & Schildhauer, 2009) and the DataONE (n.d.a) Best Practices data- base, the information on data management and curation practices is scattered across various publications, websites, and training curricula. Similarly, an environmental scan of data management and curation at Cornell University revealed that the available resources, training, and services on data management at Cornell are scattered (Block et al., 2010). Hence, Cornell formed the Research Data Management Service Group in 2010 to be "a collaborative, campus-wide organization that links Cornell University faculty, staff and students with data management services to meet their research needs" (Research Data Management Service Group, n.d., "Mission"). In the area of formal graduate student training, our scan found that several workshops and classes are available that cover various components of data management and it are conceivable that pieces of the process may be addressed in research methods classes and research labs. For example, in the Department of Natural Resources at Cornell, there are courses that cover basic biological statistics, wildlife population analysis, hydrologic data and tools, data collection and analysis for forest and stream ecology, and spatial statistics. Other departments across the College of Agriculture and Life Sciences have courses that address geographic information systems (GIS), remote sensing, spatial modeling and analysis, temporal statistics, genomics and bioinformatics. In terms of non-curricular opportunities, units such as the Cornell University Library, Cornell Statistical Consulting Unit, and Cornell Institute for Social and Economic Research offer open workshops and consultation on GIS, basic data analysis, Bayesian statistical modeling, multilevel modeling, logistic regression analysis, linear regression parameters, path analysis, mediation analysis, experimental design, longitudinal data analysis, and other statistical techniques, as well as training on GIS software packages such as ArcGIS and Manifold, and statistical software such as SAS, SPSS, Stata, and R. However, despite these opportunities, there is still a lack of comprehensive training that addresses the major elements of data management for natural resources students in a holistic fashion.

## CASE STUDY OF GRADUATE STUDENT DATA INFORMATION LITERACY NEEDS IN NATURAL RESOURCES

To discover more about data management needs at Cornell University, we used the DIL interview protocol (available for download at http://dx.doi.org/10.5703/1288284315510) to interview the faculty member in natural re- sources, one of his former graduate students, and a current lab technician during the period of March through May, 2012. Each participant rated how important DIL skills were to their data. The following section provides an over- view of the responses we received.

**Table 4.1** DIL Competency Ratings of Participants in the Natural Resources Case Study (n = 3)

| DIL Competency | Faculty Member | Former Graduate Student | Lab Technician |
|---|---|---|---|
| Discovery and acquisition of data | Somewhat important | Essential | Very important |
| Databases and data formats | Essential | Essential | Important |
| Data conversion and interoperability | Essential | Essential | Very important |
| Data processing and analysis | Essential | Essential | Very important |
| Data visualization and representation | Essential | Essential | Important |
| Data management and organization | Essential | Essential | Essential |
| Data quality and documentation | Essential | Essential | Essential |
| Metadata and data description | N/A | Essential | Important |
| Cultures of practice | Important | Essential | Essential |
| Ethics and attribution | Essential | Very important | Essential |
| Data curation and reuse | Very important | Essential | Very important |
| Data preservation | Important | Essential | Important |

The lab performed longitudinal studies of fish and zooplankton species. Some of the data sets contained information collected over decades, emphasizing the crucial need for data curation and maintenance over the extended life span of the data. Because these longitudinal data can- not be reproduced, a more formalized approach to data curation and management would be of great utility to students in the lab. The faculty member and lab staff also used databases extensively to organize and manage their longitudinal data sets. For this reason, they described acquiring the data management and organization skills necessary to work with databases and data for- mats, document data, and handle accurate data entry as essential (see Table 4.1). Otherwise, as the faculty member memorably stated, “it’s [as if] the data set doesn’t exist.”

Interviewees noted data conversion and interoperability as a particularly important skill for importing data into statistical packages. Two out of three of our respondents mentioned that they lacked an understanding of the differences between raw and processed data and how they were used. The faculty member felt that students lacked a good understanding of data visualization theory, an interesting emerging area. Less important to the faculty member was that students had an understanding of how to access external data (other than geospatial data), how to find and evaluate data repositories, and version control. The reasons varied: in some cases the faculty member felt that there was little need for the skill on that particular project; the students learned the skill informally (e.g., finding external data or data repositories through trial and error); or one or two people in the lab handled the task for everyone (e.g., entering data into Excel and the Access database).

Metadata was of high importance to all of our interviewees. When asked about metadata, the faculty member responded that he wasn’t even sure what it meant; however, he hoped to learn about it over the course of the collaboration. The former graduate student and the cur- rent lab technician placed even more emphasis on data documentation and description skills than the faculty member. The lab technician attributed much of the documentation and description he performed to a “personal coping strategy,” so that when he came back to the data later he could understand what he did and where he was in the process.

The former graduate student indicated that accessing and using external data sets, depositing data into repositories, data preservation, and intellectual property were important areas of knowledge. He learned most of what he knew through trial and error, from colleagues, and in peer-to-peer learning. Perhaps this was one of the reasons that he was adamant about best practices and training students early in graduate school. In answering our question about what he wished he’d known or been taught before graduate school, he said:

> By graduate school, that’s the point in which you are putting data in [spreadsheets], [so] your best management practices should be in place. But I recognize they’re probably not. . . . So [data management skills] should be the very first thing you learn when you come to grad school.

When asked about the importance of the DIL skills, the former graduate student listed all as essential (see Table 4.1) but noted that some were covered better than others. For example, skill development in the discovery and acquisition of data happened pretty well, but he found education about databases and data formats and data conversion and interoperability in it's in- fancy. Within certain skill sets, like data processing and analysis, the degree program included tools, techniques, and their effects on interpretation, but did not include more advanced concepts like workflow management tools. He also noted that there was a lack of norms, or weak norms, in the field regarding its cultures of practice. There was a need for those in the field, especially faculty and principal instigators (PIs) of research projects, to push for higher standards in data management issues. He felt that most of the outcomes he mentioned as essential were taught poorly or not at all.

In fact, across most of the competencies discussed, lack of formal training for acquiring important skills arose as a common theme. The student and technician noted that they acquired most of their skills informally, especially in areas such as generating visualizations and ascribing metadata to files, as there was no formal on-campus training and few readily identifiable people with expertise. Although there were classes and workshops available, students were not aware of them and were more receptive to just-in-time training or troubleshooting. When we discussed the availability of Cornell courses to learn about R, one respondent said, "I don't know if there are actual courses on it. I imagine there are somewhere, but I haven't pursued that and I don't know that I really have time to take a course." The student described the optimal situation as one where he would have access to an expert who was using R in a similar way, much like the library has a GIS librarian avail- able for GIS users.

There were some disconnects between what we learned from the faculty member and what we heard from the lab technician and the former graduate student. Discovery and acquisition of external data was only somewhat important to the faculty member. He felt that "if they didn't know these [databases] existed, it wouldn't matter," explaining that they seldom used external data in their research. However, the student and the lab technician reported using external data and exhibited limited knowledge of disciplinary repositories. Our discussion of cultures of practice skills followed the same path: it had less importance to the faculty member, but was essential to the student and the lab technician. The former graduate student's level of awareness of the skills and their necessity was very high, especially since he had had a great deal of experience as an administrator of a large data set. For ex- ample, the faculty member and the lab technician placed less emphasis on understanding formal metadata standards and data preservation (counting them as important, but not essential), in contrast to the former graduate student and what we found in the environ- mental scan and literature review. They also did not mention workflows or tools like Morpho a great deal. This disconnect between faculty and student views is unsurprising, since faculty tend to assume everyone understands the culture that they've been embedded in for years. Additionally, those who are not data- base administrators or who have not had occasion to need certain skills will naturally tend to downplay their importance.

While respondents considered nearly all of the skills we covered in our interview important, those that were not as highly prioritized included discovery and acquisition of data and data preservation.

Interestingly, there were a few differences in opinion between our faculty collaborator and the others we interviewed. The most dramatic difference was around discovery and acquisition of data, which the student and the lab technician felt was very important or essential. In contrast, our faculty collaborator felt that students should already have a good grasp of where to obtain data sets and therefore considered it only somewhat important (with the lowest rating of any of the competencies). Cultures of practice was an- other example of a competency that the faculty member felt the students should under- stand (and he rated it as “important”). This is one that the student and the lab technician felt was essential and needed to be addressed in educational interventions.

A TWOFOLD INSTRUCTIONAL APPROACH TO DATA INFORMATION LITERACY NEEDS

In fall 2012 and spring 2013 we implemented instructional interventions based on our findings to address the gaps that we found in the curriculum covering data management skills. Given the wide range of competencies of interest to the faculty and students interviewed, the Cornell DIL team narrowed the skills down ac- cording to the following principles:

1. Does the competency address a gap we found in the curriculum?

2. Did we have the expertise to address the need? If not, could we include someone else who did have the expertise?

3. Where could we add the most value?

After asking these questions in concert with our faculty collaborator, the four DIL-related areas on which we focused were data management and organization, data analysis and visualization, data sharing, and data quality and documentation. Our instructional approach was twofold: in the fall we offered workshops in the library addressing several data management topics; in the spring we offered a six-session, one-credit course for graduate students in natural resources.

*Instruction Approaches: 1-Hour Workshops and 6-Week Course*

In the fall, we offered a series of 1-hour library-sponsored workshops aimed at graduate students in the sciences, each introducing a different data management topic. The first workshop focused on data management planning and was an unqualified success: 30 students attended and we had an additional 13 on a wait list. The subsequent workshops had lower attendance: 8 attended the data organization workshop, 10 attended the data documentation workshop, and only 4 signed up for the data sharing workshop, so it was canceled. Despite the decreased attendance at the later workshops, we felt we were successful because the later session subjects were more specific, addressing topics that appealed to a more limited audience than had the broader workshop on data management (see Table 4.2). The students who at- tended were active and enthusiastic participants and expressed appreciation after the workshops.

TABLE 4.2 - Weekly Course Topics and Readings in the Spring 2013 One-Credit Cornell Course NTRES 6940: Managing Data to Facilitate Your Research

| Topic | Description and Readings |
|---|---|
| 1. Introduction to data management | We will use the first class session for introductions and logistics. The instructors will give a brief explanation of DMPs and reasons for using them. We'll then have a group discussion of research, data problems encountered, and data management needs.<br><br>Readings:<br>• Wolkovich, E. M., Regetz, J., & O'Connor, M. I. (2012). Advances in global change research require open science by individual researchers. Global Change Biology, 18(7), 2102–2110. http://dx.doi.org/10.1111/j.1365-2486.2012.02693.x<br>• National Science Foundation (n.d.). Dissemination and sharing of research results. http://www.nsf.gov/bfa/dias/policy/dmp.jsp<br>• Research Data Management Service Group (n.d.). Data management planning: Guide to writing a data management plan (DMP). http://data.research.cornell.edu/content/data-management-planning |
| 2. Data organization | Organizing your data at the front end of a research project will save time and increase your ability to analyze data. This session will introduce you to the principles involved in creating a relational database and will provide examples to help you organize your own data in this manner. Topics will include best practices for data entry, data types, how to handle missing data, organization by data type, and data file formats.<br><br>Readings:<br>• Borer, E. T., Seabloom, E. W., Jones, M. B., & Schildhauer, M. (2009). Some simple guidelines for effective data management. Bulletin of the Ecological Society of America, 90(2), 205–214. http://dx.doi.org/10.1890/0012-9623-90.2.205<br>• Research Data Management Service Group (n.d.). Preparing tabular data for description and archiving. http://data.research.cornell.edu/content/tabular-data<br>• |
| 3. Data analysis and visualization | Analyze existing data and create graphs using R in order to effectively communicate findings.<br><br>Readings:<br>• DataONE (n.d.). Education modules: Lesson 10—Analysis and workflows. http://www.dataone.org/education-modules<br>• Noble, W. S. (2009). A quick guide to organizing computational biology projects. PLoS Computational Biology, 5(7), e1000424. http://dx.doi.org/10.1371/journal.pcbi.1000424 |
| 4. Data sharing | The NSF and other funding agencies have already adopted data sharing policies. Publishers also have data sharing requirements, whether they host data themselves, or expect researchers to deposit data in a data center or to make it available upon request. So where to share? During this class session, we'll discuss disciplinary databases, Cornell's eCommons digital repository, and some other sharing strategies, and will talk about evaluation criteria upon which to base your decision about where |

| | |
|---|---|
| | to share your data.<br><br>Readings:<br>• Center for Research Libraries. (2005). General factors to consider in evaluating digital repositories. Focus on Global Resources, 25(2). http://www.crl.edu/focus/article/486Databib \| searchable catalog of research data repositories (http://databib.org/index.php)<br>• eCommons@Cornell (http://ecommons.library.cornell.edu/) |
| 5. Data quality and documentation | While written documentation—for example, in a lab notebook—is still important, the platforms on which modern researchers are working and collecting data are increasingly complex. How do you document your digital data and the steps you take to analyze it? Are your files sufficiently organized and well described so that others can interpret what you've done? What about yourself, 3 months from now? During this class session on data documentation, we'll discuss the challenge of remembering details relevant to interpreting your data, and offer some best practices and strategies to adopt in order to organize and describe your data for yourself and others.<br><br>Readings:<br>• Disciplinary Metadata \| Digital Curation Centre (http://www.dcc.ac.uk/resources/metadata-standards)<br>• Kozlowski, W. (2014). Guidelines for basic "readme" style scientific metadata. http://data.research.cornell.edu/sites/default/files/SciMD_ReadMe_Guidelines_v4_1_0.pdf<br>• Rudstam L. G., Luckey, F., & Koops, M. (2012). Water quality in offshore Lake Ontario during intensive sampling years 2003 and 2008: Results from the LOLA (Lake Ontario Lower Foodweb Assessment) Program. http://hdl.handle.net/1813/29691 |
| 6. Final wrap-up: data management plans | For the final class session, participants will have the opportunity to present a DMP for peer discussion and review. Depending on interest, presentations may range from 6 to 15 minutes.<br><br>Readings:<br>• Sample DMP from Inter-University Consortium for Political and Social Research (ICSPR) (http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/dmp/plan.html)<br>• Sample DMPs from University of California San Diego (http://idi.ucsd.edu/data-curation/examples.html)<br>• Sample DMPs from the University of New Mexico (http://libguides.unm.edu/content.php?pid=137795&sid=1422879) |

In the spring, the Cornell DIL team offered the six-session, one-credit course for graduate students in natural resources, Managing Data to Facilitate Your Research. The data librarian and the faculty collaborator co-taught the course. The content was similar to the fall semester library workshops, but we were able to build on prior classes as we progressed through the material. For example, in the workshop for- mat, we introduced the basics of data management as part of each workshop; in the course format we introduced data management in the first session and were focused on additional content in each subsequent class. At the beginning of each session, we recapped what we covered in the last session and offered time to respond to questions. Because we listed the course through the Department of Natural Resources, we had a more subject-specific focus and drew on examples from ecology and fisheries research. For example, during the session on data analysis and visualization, the faculty collaborator demonstrated linking stable isotope data from the Cornell University Stable Isotope Laboratory to the master database file from the Adirondack Fisheries Research Program. This involved discussing data import, linking the new table to master tables in the database, developing a query, and exporting the data into Microsoft Excel. All of these topics could have been discussed without the context of real re- search data, but using real-life examples drawn from the discipline helped the students under- stand what was happening in the data management process and, more importantly, why it should happen. We created a library guide for the course, available at: http://guides.library.cornell.edu/ntres6940.

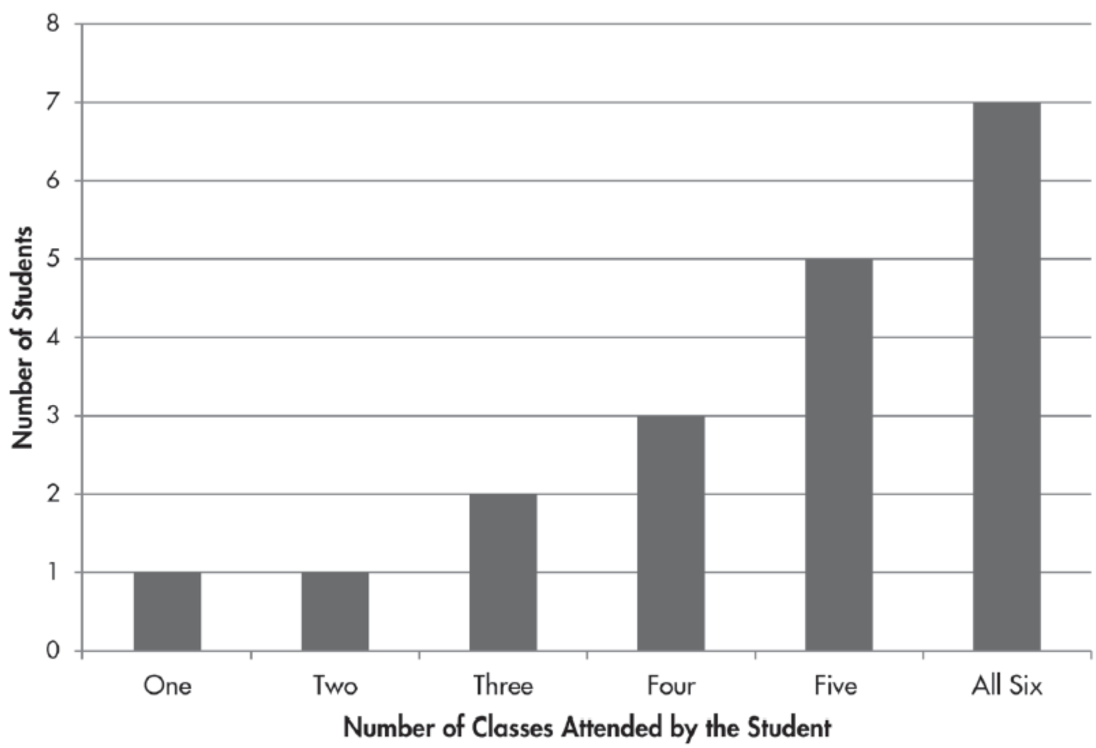


**Figure 4.1** Self-reported attendance by students enrolled in the spring 2013 one-credit Cornell course NTRES 6940: Managing Data to Facilitate Your Research (n = 19).

We drew on several resources to build the course and workshop content. For example, DataONE (n.d.b) created education modules covering data management topics that are openly available at http://www.dataone.org/education-modules. We relied heavily on those that matched our identified needs. We did make changes to the slides, adjusting for the discipline and for the time allotted. We also made use of an Ecological Society of America (ESA) publication about best practices in data management (Borer et al., 2009). Twenty-five students enrolled in the course. Most of the students were from the natural resources department, though there were students from biological and environmental engineering, ecology and evolutionary biology, crop and soil sciences, and civil and environ- mental engineering. The students ranged from first-year to fourth-year graduate students. Two faculty and staff members attended. Fifteen students attended four or more of the six sessions in the course (see Figure 4.1). Given that it was only a one-credit, 6-week-long course, we could only briefly touch upon the major issues. A mix of higher level, conceptual articles gave context to our discussions, along with more practical resources for students to explore on their own and pointers to Cornell University resources for training and just-in-time help.

## LEARNING OBJECTVIVES FOR THE CORNELL COURSE

The aim of our instruction for the course was to introduce students to data management best practices in natural resources and to help students create plans to manage their data effectively and efficiently while meeting funder and publisher requirements. The learning objectives were as follows. By the end of this course, students will be able to:

- describe data management and why it is important;
- describe their research and data collection process in order to identify their data life cycle and complete the initial part of the DMP;
- evaluate a DMP to recognize the necessary components of a successful plan;
- describe and follow best practices in structuring relational databases to make analysis and retrieval easier/more efficient for long-term studies;
- analyze existing field data and create graphs using R to effectively communicate findings;
- evaluate disciplinary data repositories to determine requirements and fitness for data deposit;
- evaluate the annotation/documentation accompanying a data set to recognize the appropriate level necessary for long-term understanding by self and others;
- create a DMP to manage and curate their own data for effective long-term use and reuse as well as to meet funding requirements.

Each session attempted to meet the learning outcomes outlined by the DIL project (see Table 4.3). We addressed them through a variety of activities; however, we were not able to address all of them in great depth. Some sections of the course were more traditional. For example, students read an article on effective data management practices (Borer et al., 2009) before class and commented to a discussion forum on points they found interesting or that needed more clarification. Then we reviewed the comments and discussed them in class.

We considered graduate students to be expert learners; therefore we employed collaborative learning techniques, including think- pair-share and group problem solving (Center for Teaching Excellence, 2013b). For example, as a class activity students discussed their research data life cycle in detail and then drew a diagram of the stages of research. For "evaluate disciplinary data repositories" students worked in groups to identify possible repositories for data deposit for their subject. (See Appendix B to this chapter for a full description of the exercise.) For the session on data documentation, students worked in groups with examples of metadata and evaluated what was done well and what could be improved. Finally, we asked those who chose to complete the optional DMP exercise to complete a different section of the DMP each week, and participants received feedback from the librarian instructors.

TABLE 4.3 - Needs and Learning Outcomes Addressed in the Cornell For-Credit Class per Session

| Session | Needs Identified | Outcomes Addressed |
| --- | --- | --- |
| 1. Introduction to data management | Basic introduction to data management: importance in the research context of the audience | Understands the life cycle of data, develops DMPs, and keeps track of the relation of subsets or processed data to the original data sets<br>Creates standard operating procedures for data management and documentation |
| 2. Data organization | Acquiring the data management and organization skills necessary to work with databases and data formats, document data, and handle accurate data entry is described as essential, otherwise, "it's as if the data set doesn't exist" | Understands the concept of relational databases, how to query those databases, and becomes familiar with standard data formats and types for discipline<br>Understands which formats and data types are appropriate for different research questions |
| 3. Data analysis and visualization | A good understanding of higher end data visualization, though not positioned as currently essential but as an interesting emerging area by the instructor, was in short supply. The lab primarily uses R for data analysis and visualization, but training is limited, and not aimed specifically at students in natural resources | Becomes familiar with the basic analysis tools of the discipline<br>Uses appropriate workflow management tools to automate repetitive analysis of data<br>Proficiently uses basic visualization tools of discipline |
| 4. Data sharing | Areas such as accessing external data (except for background geospatial data) and finding and evaluating data repositories were of less importance to the faculty member than to the | Recognizes that data may have value beyond the original purpose, to validate research or for use by others<br>Locates and utilizes disciplinary data repositories |

| | | |
|---|---|---|
| | students, but the faculty member expressed interest in learning more about Cornell's institutional repository | |
| 5. Data quality and documentation | Skills such as ascribing metadata to files are acquired informally; furthermore, the faculty member noted he wasn't even sure what was meant by metadata, and he hoped to learn about it over the course of the collaboration | Recognizes that data may have value beyond the original purpose, to validate research or for use by others<br>Understands the rationale for metadata and proficiently annotates and describes data so it can be understood and used by self and others<br>Develops the ability to read and interpret metadata from external disciplinary sources<br>Understands the structure and purpose of ontologies in facilitating better sharing of data |
| 6. Data management plans | Funders and other organizations are increasingly requiring DMPs, and few graduate students are aware of the components of a good DMP | Understands the life cycle of data, develops DMPs, and keeps track of the relation of subsets or processed data to the original data sets<br>Creates standard operating procedures for data management and documentation<br>Articulates the planning and actions needed to enable data curation |

ASSESSMENT

We used a combination of formative and summative assessment tools, including 1-minute reflections after each session, feedback on outputs from active learning exercises, and a final survey (Center for Teaching Excellence, 2013a; Downey, Ramin, & Byerly, 2008). A 1-minute reflection was administered either as a survey after each library workshop, or as a discussion question via the course Blackboard site. Figure 4.2 shows a typical 1-minute reflection assignment.

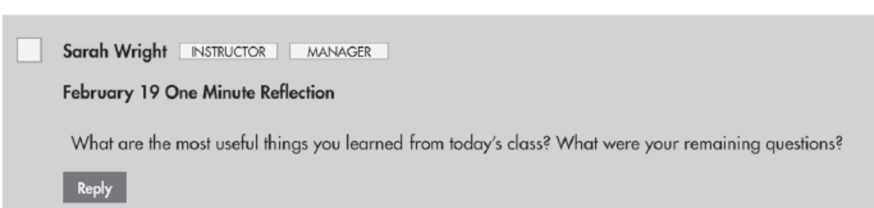


**Figure 4.2** One Minute Reflection assignment via a course Blackboard site.

In addition to the 1-minute reflection posts, we used the discussion board for students to ask questions after each class session. There were 68 posts, with 21 participants—representing the majority of the students enrolled. We gained many substantial and useful comments using this method. In fact, the comments were so useful that it became our practice to review the most pertinent comments at the beginning of each class as a way to emphasize content from the last class or to lead into content for that day's class. After the class on data organization and the use of relational databases, we received positive feedback from students enthusiastically discussing the changes they would make due to what they had just learned.

> The "rules of thumb" were a great summary of various best practices for data management. It was interesting to read that computer code was actually a form of metadata in itself. I suppose I had never looked at it in that light before but from now on I will take my commenting more seriously! I was also grateful for the explanation of best practices for relational databases. I've heard of the term but this paper did a great job walking through the formation of one, step by step. Finally, I'm finding that by taking this class and doing these readings
>
> I'm becoming more aware of different data management services in my own field.
>
> Three points from Borer et al. (2009) that were particularly useful: [1] the merits of using scripted analyses. Having used JMP for 4 years, I know too well the agony of trying to replicate drop-down menu instructions months after doing an analysis. I plan to switch to R. [2] standardized file naming system using the international date format. While I use descriptive folder names, I do not always use descriptive file names and I am not consistent with date format . . . [which] makes searching for files on my computer inefficient . . . [and] also means that when I send others my data it loses some descriptive information . . . [3] full taxonomic names in data files. A few years ago I did an experiment in which I identified 100+ plant species in the field. I used abbreviations in my data. Flash forward 3 years, and it took me days to reconstruct what all my abbreviations were. Some taxonomic names had changed. Never again!

We received comments that required follow- up and more conversation:

> The relational database method seems great but will take some getting used to. Is there a way to connect excel and access files that would allow you to input data and automatically update in the relational database?
>
> Learning about relational databases was very useful. Efficient organization of spreadsheets was also helpful. I would like to learn more about how to organize metadata, but I think this is an upcoming class discussion. Also, I am still lacking clear reasons why Access is preferable to Excel. What does Access offer that Excel does not? What are the features that make Access particularly useful?

After reading the last comment, we felt that we had not clearly explained the advantages of a relational database, so we addressed that point at the beginning of the next class. As these examples illustrate, the

1-minute reflections proved to be a powerful form of formative assessment that allowed us to respond to the learning needs of the students.

We also provided active learning exercises so that students could receive outcomes-based assessments of their work and understanding. Some of these were in-class exercises that we collected and delivered feedback on for the students. Others were optional out-of-class assignments, which included rubrics for assessment. Though few students completed the optional assignment (n = 5), all who tried it found it useful; those who didn’t complete it indicated that it probably should be required in the class. In most cases, we simply discussed what students found during the exercises and gave feed- back during discussion.

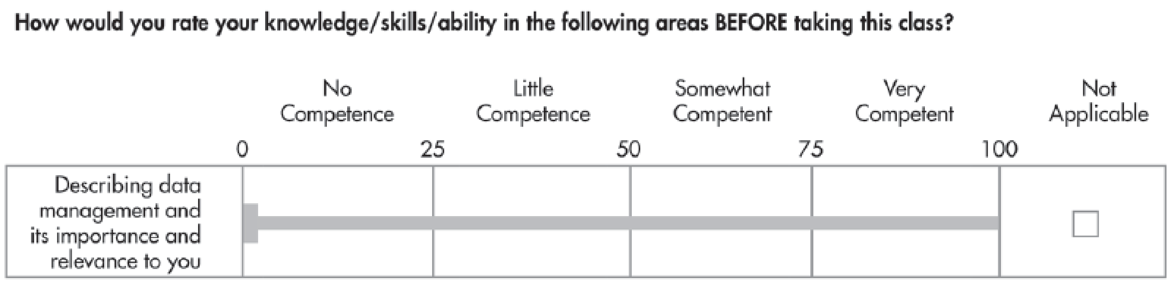


**Figure 4.3** Example self-assessment survey question using a slider scale.

Finally, we administered a self-assessment survey at the end of the class to gauge the success of our experimental course (see the full instrument in Appendix C to this chapter). We invited and received constructive criticism via the survey instrument, much of which will guide our next attempt at offering similar instruction. Here, we also asked the students to self- evaluate their skill levels concerning the course outcomes both before and after taking the class (see an example in Figure 4.3). Rather than performing pre- and post-evaluations, we asked students to rate their skill levels before and after, after instruction occurred. This method avoided the problem of overestimation of skill that is common before learning a new topic (Kruger & Dunning, 1999). Having learned more about the course outcomes, students could then better compare what they actually knew at the beginning to what they had learned during the course.

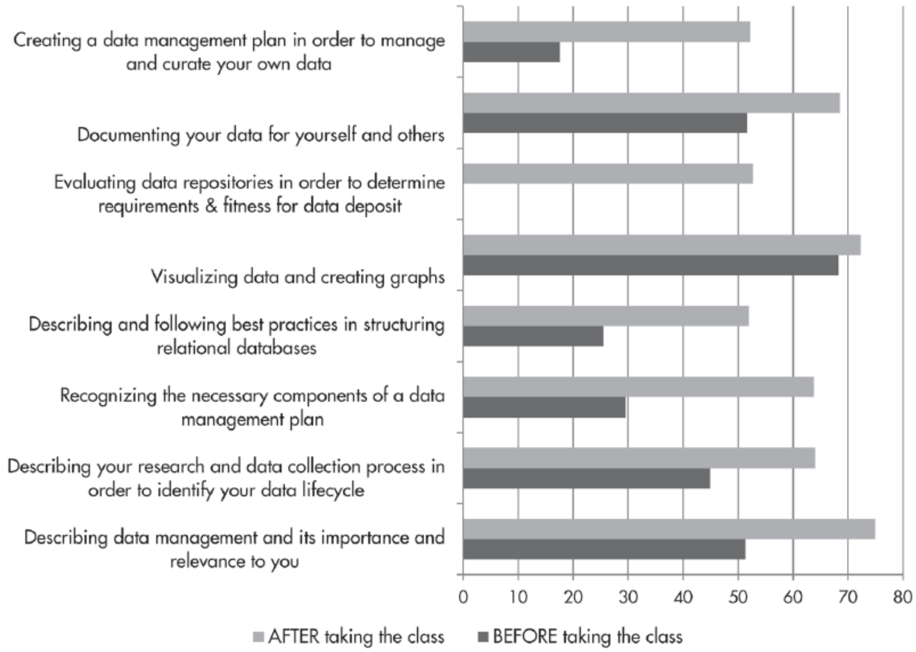


**Figure 4.4** Average responses (before and after) to the survey question, "How would you rate your knowledge/skills/ability in the following areas?" (n = 17). (NOTE: Due to a technical error in the survey, the student response to "Evaluating data repositories in order to determine requirement & fitness for data deposit" before taking the class could not be included in this figure.)

On average, responses (n = 17) showed marked increases in the skills, knowledge, and abilities that the students felt they possessed after taking the class, as shown in Figure 4.4. However, there was room for improvement since on average students rated none of the outcomes in the "somewhat competent" to "very competent" range after the course. In fact, several outcomes received an average rating of "little competence" and "somewhat competent" following the course. And, the most frequently voiced criticism of the class was that we touched on a lot of important topics, but we didn't have time to go in-depth and failed to provide enough opportunities to practice what we'd discussed. Still, feedback was overwhelmingly positive, and the majority of students (13 out of 16) would recommend this course to others in natural resources.

RESULTS

Overall, the response to the workshops and the course was very positive. Students reported a better awareness of data management skills and the resources and tools available to them. One student noted, "I think the topic of this class is SO ESSENTIAL [to] the way scientific re- search is being carried out and shared now. . . . [This course] fills a hole in Cornell grad education." Filling a need in the curriculum is exactly what the Cornell DIL team was trying to do, and it was gratifying that students recognized the importance of the topic and appreciated our educational efforts!

The self-reported increase in skill for all of the learning outcomes was another positive outcome of the course. The marked increase in students' abilities to articulate the importance of data management, to create their own DMP, and to de- scribe and document their own data collection practices was an important step forward. Their comments in the end-of-class survey bore this out and indicated their increased awareness of many areas of data management. As one student said:

> I think just the exposure to the different aspects of data management and the discussion about the usefulness of relational databases and analysis software like R can be of great benefit to students, especially those that are relatively new to research and may not be aware of the types and benefits of resources available to them.

Benefits for the DIL project included uncovering areas in which there was a need for more exploration, such as curation of training resources and opportunities, direct instruction on tools (e.g., conversion from Excel to database pro- grams, database tools for Mac users, data visualization tools, qualitative analysis tools like Atlas. ti), and allowing students to exchange information and network with each other. Interestingly, in the final class students exchanged information about ad hoc training in data visualization in departments beyond natural resources. This shows the library's potential role in facilitating peer-to-peer training in addition to the formal, instructor-led educational initiatives. The library is experimenting with the role of facilitator to crowd source tips and workflows from students who have expertise and to schedule project clinics with interested and skilled students and staff. This facilitator role could be fruitfully applied to DIL and would address the need to balance a great need for specialized instruction with a small library staff that has limited time and skills.

Before the course ended, the project team at Cornell discussed how to continue providing data management instruction and what could be done to improve it. This project has been an exciting experiment, but there is much interest beyond the library. Our faculty collaborator discussed how to offer this course next time—indicating even before we had finished the course that he was invested in doing it again. Building a stronger relationship with this faculty member and investigating the students' need for hands-on training (in areas that faculty assumed the students knew or would learn informally along the way) was one of the most rewarding parts of this experience.

The course also gained wider recognition among faculty and students; it was the focus of a short article in the Cornell Chronicle titled "Course Teaches Grad Students How to Man- age Their Data," which sparked inquiries from faculty and graduate students in other departments (Glazer, 2013). This prompted the library to hold more one-time sessions and to add modules to online guides that hopefully will lead to more course- and curriculum- integrated instruction.

Although the student feedback was very positive, there is room for improvement. For example, the scope of the course should be more focused, and it would work better with a smaller group that has a similar level of experience. We would like to expand the course beyond six sessions, or eliminate content if we are unable to increase the number of sessions. In the current course, we included more material than we could reasonably cover. These changes would also allow us to introduce more exercises and to provide more opportunities for hands-on learning. This was a major criticism received of the course. Including more practical exercises in the course and holding project clinics and peer-led workshops would provide students the opportunity to experiment with and learn using their own research data. These formats would also allow students more time for discussion and peer exchange around personal workflows and existing practices. They would make the sessions less prescriptive and instructor-led and more student-led and free-flowing. Discussions would also allow for more just-in-time exchange of information for students who are interested in particular areas, and for more advanced students who might not want to take a full course.

With these goals in mind, we plan to pro- vide general, beginner-level data management library workshops in the fall, open to anyone, focused on topics like creating a DMP or writing a readme file to describe your data. We'll then provide a disciplinary course (possibly in other departments that have expressed interest) where we can provide more focused, in-depth instruction and require active learning components, such as the creation of a DMP. The peer- to-peer workshop model and project clinics are also a possibility for the future.

It is clear that DIL skills are important skills that graduate students feel are not being taught sufficiently in their programs. A former graduate student brought up the need for data management instruction even earlier, stating, "I think it starts as an undergraduate. It's an easily understood discipline at even a high school or junior high level, and I would start it that early, if possible." We would like to incorporate data management instruction into undergraduate laboratory classes, similar to the way we've incorporated information literacy into the curriculum at multiple points in programs. This is a long-term goal that has grown out of the current project, and it will require collaboration and the investment of groups both inside and outside the library.

## DISCUSSION

The Cornell DIL team entered this project with a general idea of the DIL competencies; however, interacting with students and teaching the competencies resulted in some changes to our original impressions. Much as the ACRL's (2000) information literacy competency standards outline high-level outcomes for information literacy across an entire curriculum, the DIL competencies are a starting point for articulating what data management concepts students should understand and apply throughout their research careers. How this plays out at varying stages of a researcher's education and for each discipline is a much more detailed and idiosyncratic issue. We found that many of the students in the course, especially those at the beginning skill levels in a particular competency, wanted much more specific (and often very tool-based) skills (e.g., how to better use spreadsheet and database packages like Excel and Access), rather than the higher level conceptual DIL skills, especially in the absence of an immediate real-world application (e.g., funder data sharing requirements).

Since the competencies outlined in the DIL project covered such a wide range in a quickly changing field, they placed an emphasis on the recognition and understanding of general best practices and much less emphasis on the skills needed at the disciplinary and lab/project level. Working with the general DIL competencies and tailoring them to course and class session outcomes forced us to refine and articulate what we wanted students to be able to do and how we wanted them to demonstrate and apply their understanding to their disciplinary-specific situation. For example, we recognized that skills build in a progression, so we derived the following outcomes from the general DIL competency "understands the life cycle of data, develops DMPs, and keeps track of the relation of subsets or processed data to the original data sets":

- Describe research and data collection process to identify data life cycle and complete initial part of DMP
- Evaluate a DMP to recognize the necessary components of a successful one
- Create a DMP to manage and curate own data for effective long-term use and reuse as well as to meet funding requirements

In the course, we briefly addressed tracking subsets of data, but addressing this topic alone was much more involved than it first appeared. This pattern emerged in working with the competencies.

The range of skill levels in the class and the wide variety of types of data with which they worked (e.g., quantitative and qualitative; small and large data sets in multiple formats) showed the need for competencies that progressed over time from basic understanding and tool-based skills to higher level competencies in analysis and synthesis, as well as for outcomes that ad- dressed particular disciplines or kinds of data. This work is the beginning of that effort.

Questions we asked ourselves in the process of creating the workshop series and the for- credit course map well to areas that we need to address to integrate DIL competencies into the curriculum:

- What skills do students currently have and where are their most pressing needs? The interviews we conducted with the faculty member and students in natural resources gave us an in-depth view of the skills and attitudes of a very small sample. A larger survey of graduate students and faculty in natural resources (and other disciplines) would give a better idea of the needs of the campus community.
- What are the gaps in the curriculum?
- What outcomes are already addressed, where, and at what levels? As part of the environmental scan, we identified the training available, but a closer look at the syllabi of courses that incorporated DIL outcomes and a census of available workshops and other training could help us target our efforts.
- Do we have the expertise to address student and researcher needs? If not, could we include someone else in or provide staff professional development to gain the missing expertise? It does no good to plan instruction if we do not have the expertise to de- liver it, so we asked ourselves: Who is the best person to answer this need?
- Where can we add the most value? Where can we find partners to supplement areas that are

outside our purview? Strategic partnerships with other departments on campus can help reach students at the time of need.

- What curriculum resources already exist to meet particular DIL outcomes and at what level? Instead of reinventing the wheel, we should try to find, centralize, and adapt available curriculum resources for DIL educational content. A repository or di- rectory of curriculum resources for DIL would be useful.

## CONCLUSION

We are only beginning to specify the competencies in DIL that will develop the data management skills that future researchers and scientists will need, and many barriers to identifying them still exist. The rapidly changing nature of the field, the heterogeneity of skills within the disciplines, and the intensive and long-term nature of the task of integrating DIL skills within (and alongside) the curriculum present challenges to academic librarians seeking to take on this task. The questions posed in our discussion are a start. Similarly, the workshop series and for-credit course that we piloted at Cornell University are just a beginning. And the harsh reality is that it is impossible to scale or sustain workshops or credit courses to reach graduate students in all disciplines. These interventions may work best as gateways to introduce students to the range of skills they need to acquire through other more targeted workshops and classes, throughout their academic career. However, by taking the lessons learned in these preliminary initiatives, and by using the modules we created or adapted, we can build on this foundation to create an integrated, progressive DIL program that will prepare students for the challenges and changes ahead.

## NOTE

This case study is available online at http://dx.doi.org.10.5703/1288284315476.

## REFERENCES

Association of College and Research Libraries (ACRL). (2000). Information literacy competency standards for higher education. Retrieved from http://www.ala.org/acrl/sites/ala.org.acrl/files/content/standards/standards.pdf

Barros, E., Laender, A., Gonçalves, M., Cota, R., & Barbosa, F. (2007). Transitioning from the eco- logical fieldwork to an online repository: A digital library solution and evaluation. International Journal on Digital Libraries, 7(1–2), 109–112. http://dx.doi.org/10.1007/s00799-007-0024-7

Block, W. C., Chen, E., Cordes, J., Dietrich, D., Krafft, D. B., Kramer, S., . . . Steinhart, G. (2010). Meeting funders' data policies: Blueprint for a research data management service group (RDMSG). Retrieved from Cornell University Library eCommons website: http://hdl.handle.net/1813/28570

Borer, E. T., Seabloom, E. W., Jones, M. B., & Schildhauer, M. (2009). Some simple guidelines for effective data management. Bulletin of the Ecological Society of America, 90(2), 205–214. http://dx.doi.org/10.1890/0012-9623-90.2.205

Carr, M. H., Woodson, C. B., Cheriton, O. M.,Malone, D., McManus, M. A., & Raimondi, P. T. (2011). Knowledge through partnerships: Integrating marine protected area monitoring and ocean observing systems. Frontiers in Ecology and the Environment, 9(6), 342–350. http://dx.doi.org/10.1890/090096

Center for Teaching Excellence. (2013a). Active learning. Retrieved from http://www.cte.cornell.edu/teaching-ideas/engaging-students/active-learning.html

Center for Teaching Excellence. (2013b). Collaborative learning. Retrieved from http://www.cte.cornell.edu/teaching-ideas/engaging-students/collaborative-learning.html

DataONE. (n.d.a). Best practices. Retrieved from https://www.dataone.org/best-practices

DataONE. (n.d.b). Education modules. Retrieved from http://www.dataone.org/education

Downey, A., Ramin, L., & Byerly, G. (2008). Simple ways to add active learning to your library instruction. Texas Library Journal, 84(2), 52–54.

Fegraus, E. H., Andelman, S., Jones, M. B., & Schildhauer, M. (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(3), 158–168. http://dx.doi.org/10.1890/0012-9623(2005)86[158:MTVOE D]2.0.CO;2

Gil, I. S., Hutchison, V., Frame, M., & Palanisamy, G. (2010). Metadata activities in biology. Journal of Library Metadata, 10(2–3), 99–118. http:// dx.doi.org/10.1080/19386389.2010.506389

Glazer, G. (2013, Feb. 21). Course teaches grad students how to manage their data. Cornell Chronicle. Retrieved from http://www.news.cornell.edu/stories/2013/02/course-teaches-grad-students-how-manage-their-data

Global Lake Ecological Observatory Network (n.d). Understand, predict and communicate the role and response of lakes in a changing global environment. Retrieved from http://www.gleon.org/

Jones, M. B., Schildhauer, M. P., Reichman, O. J., & Bowers, S. (2006). The new bioinformatics: Integrating ecological data from the gene to the biosphere. Annual Review of Ecology Evolution and Systematics, 37(1), 519–544. http://dx.doi.org/10.1146/annurev.ecolsys.37.091305.110031

Knowledge Network for Biocomplexity. (n.d.). Knowledge network for biocomplexity. National Science Foundation. Retrieved from https://knb.ecoinformatics.org/

Knowledge Network for Biocomplexity. (2005). The knowledge network for biocomplexity (KNB) and our graduate student research training seminar series. Retrieved February, 15, 2013, from https://knb.ecoinformatics.org

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self- assessments. Journal of Personality and Social Psychology, 77(6), 1121–1134. http://dx.doi.org/10.1037/0022-3514.77.6.1121

Long Term Ecological Research Network, The. (2013). The LTER network. Retrieved from http://www.lternet.edu/

Magnusson, A., & Hilborn, R. (2007). What makes fisheries data informative? Fish and Fisheries, 8(4), 337–358. http://dx.doi.org/10.1111/j.1467-2979.2007.00258.x

McKiernan, G. (2004). Ecological and environ- mental data. Science & Technology Libraries, 23(4), 95–104. http://dx.doi.org/10.1300/J122 v23n04_08

Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B., & Stafford, S. G. (1997). Nongeospatial metadata for the ecological sciences. Eco- logical Applications, 7(1), 330–342. http://dx.doi.org/10.1890/1051-0761(1997)007[0330:NM FTES]2.0.CO;2

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: Supporting ecology as a data-intensive science. Trends in Ecology & Evolution, 27(2), 85–93. http://dx.doi.org/10.1016/j.tree.2011.11.016

Michener, W. K., Porter, J., Servilla, M., & Vanderbilt, K. (2011). Long term ecological research and information management. Ecological Informatics, 6(1), 13–24. http://dx.doi.org/10.1016/j.ecoinf.2010.11.005

Research Data Management Service Group. (n.d.). About the Research Data Management Service Group. Retrieved from http://data.research.cornell.edu/content/about

Wolkovich, E. M., Regetz, J., & O’Connor, M. I. (2012). Advances in global change research re- quire open science by individual researchers. Global Change Biology, 18(7), 2102–2110. http://dx.doi.org/10.1111/j.1365-2486.2012.02693.x