ISEES-WSSI Lessons for Sustainable Science Software from an Early Career Training Institute on Open Science Synthesis

Lenhardt, W. C.[1]; Ahalt, S.[1]; Jones, M.[2]; Aukema, J.[3]; Hampton, S.[4]; Idaszak, R.[1]; Rebich-Hespanh, S.[2]; Schildhauer, M.[2]

Introduction

Recent emphasis on software elements of cyberinfrastructure in the context of domain science has raised awareness of the importance of sustainable software for research. This emphasis also highlights the need for training and career development in these areas. This paper presents a brief summary of these challenges and describes a relatively new and unique training approach developed to address these challenges. This effort, the Open Science for Synthesis (OSS) Training Institute, a joint partnership of two NSF-funded projects-- the Institute for Sustainable Earth and Environmental Software (ISEES, http://isees.nceas.ucsb.edu) and the Water Science Software Institute (WSSI, http://waters2i2.org)-- was conducted during the summer of 2014, bringing together approximately 40 early career researchers in a three week intensive bicoastal learning experience. This paper will describe the approach and present some early lessons learned and outline a few preliminary assessment results.

Description of training need

We argue in this paper that a key element in promoting sustainable software is training. However, the training need is more comprehensive than just teaching software engineering and software development skills to researchers. The need is for broader training with software sustainability as a major element. Broader training addresses a number of obstacles. The first obstacle, the widespread lack of capacity among researchers in the environmental sciences to conduct data-intensive science, stems from a lack of relevant foundational skills. Foundational skills in data management, programming, and other fundamentals of computer science are not consistently taught in environmental sciences curricula (Hernandez et al. 2012, Strasser & Hampton 2012). Furthermore, increasing data and data complexity also drive the need for training. For example, emerging ecological observatories worldwide are creating new environmental data sets (Hamilton et al. 2007, Keller et al. 2008, Baptista et al. 2008). Similarly, another driver is the increasing capacity for remote sensing tools to contribute to environmental research (Pfeiffer et al. 2012). While these technologies hold great promise for environmental research, researchers themselves still know relatively little about how to access and use these products (Newton et al. 2009, Hernandez et al. 2012). A recent survey of graduate students in the environmental sciences (Hernandez et al. 2012) illustrates this phenomenon. Over 80% of students had received no formal training in computing or informatics at even the most basic level, and 74% stated that they had no skills in any programming language. While 72% of the students said they understood the term 'metadata,' about half had not created metadata for their dissertation data and had no plans to do so.

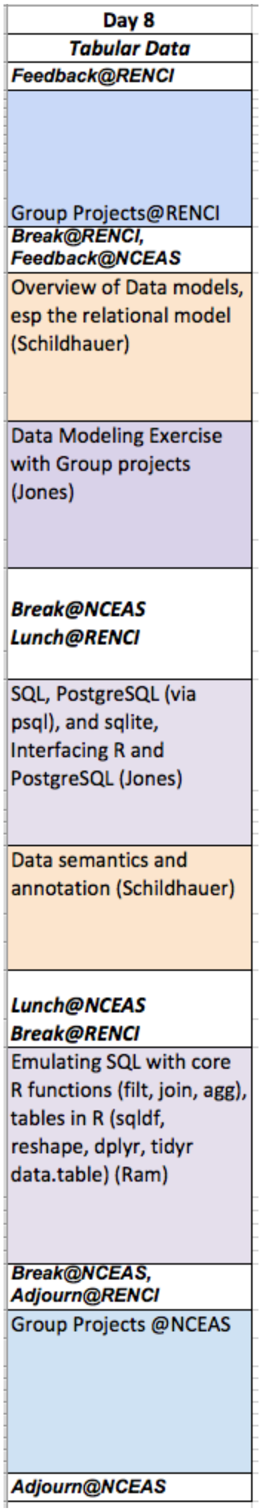
[1] Renaissance Computing Institute (RENCI)
[2] National Center for Ecological Analysis and Synthesis (NCEAS)
[3] U.S. Agency for International Development (USAID)
[4] Center for Environmental Research, Education and Outreach, Washington State University

Deeply connected to the data obstacles are software challenges. Software is integral to the scientific process in all phases. The challenges are well documented and multifaceted. Science software challenges include the wide array of software used and/or created by scientists in the conduct of their research from scripts to models that run on HPC; code documentation and reuse; lack of understanding of basic software development and engineering best practices; software development costs; code maintenance; as well as credit and attribution (Ahalt, et al. 2014). As has been reported elsewhere, scientists' training and understanding of the importance of good software practices is lagging while at the same time the reliance on developing software code is increasing (Merali 2010, Hannay, et al. 2009). In this context the issue of science teams and cross-domain expertise is also relevant. The question becomes whether to train scientists to be good coders, or to train coders in domain specific knowledge (Katz, et al 2014). Finally, and not surprisingly, there are the natural connections between software and data in the context of science. Data and code are intertwined (Lenhardt, et al 2014) as are the issues related to training and sustainability.

Description of the training approach

ISEES and WSSI teamed to implement an intensive three-week training institute to address the challenges and deficiencies described above. The institute, Open Science for Synthesis[5] (OSS), combines intensive, in-person training in advanced research skills with active collaborative research on critical topics related to ecological and environmental science. A unique bi-coastal training opportunity, OSS is designed to give early career scientists new data and computational skills to support open, collaborative, and reproducible synthesis research. The 2014 training institute was based on a successful version run solely by NCEAS in 2013.

During the institute participants received hands-on guided experience using best practices in the technical aspects that underlie successful open science and synthesis – including data discovery and integration, analysis and visualization, as well as special techniques for collaborative scientific research, including virtual collaboration over the Internet. The training addressed the following themes: collaboration modes and technologies, virtual collaboration; data management, preservation, and sharing; data manipulation, integration, and exploration; scientific workflows and reproducible research; agile and sustainable software practices; data analysis and modeling; as well as communicating results to broad communities. The institute is structured to leverage instruction, discussions, and hands-on exercises. A key component is a real-time synthetic scientific research process. This is accomplished via daily work on group synthesis projects. A version of the full schedule may be accessed at http://bit.ly/OSS014Schedule.



Day 8
Tabular Data
Feedback@RENCI
Group Projects@RENCI
Break@RENCI, Feedback@NCEAS
Overview of Data models, esp the relational model (Schildhauer)
Data Modeling Exercise with Group projects (Jones)
Break@NCEAS Lunch@RENCI
SQL, PostgreSQL (via psql), and sqlite, Interfacing R and PostgreSQL (Jones)
Data semantics and annotation (Schildhauer)
Lunch@NCEAS Break@RENCI
Emulating SQL with core R functions (filt, join, agg), tables in R (sqldf, reshape, dplyr, tidyr data.table) (Ram)
Break@NCEAS, Adjourn@RENCI
Group Projects @NCEAS
Adjourn@NCEAS

Figure 1: Typical Training Institute Daily Schedule

[5] https://www.nceas.ucsb.edu/OSS

The collaboration theme for the training institute focuses on sensitizing the participants to some of the sociological and process issues related to group research collaboration. These issues include data sharing, authorship attribution, and group dynamics. The exercises are designed to highlight issues better addressed early in a collaborative project and dealt with in a straightforward manner. Participants are provided with examples of data sharing and attribution agreements and are encouraged to apply the documents in the context of the group projects. The training itself demonstrated the application of these approaches and technologies.

The data management and software development best practices themes covered a wide range of topics from data management best practices, database technologies, data modeling, software engineering and design. The abstract lessons were reinforced with exercises and hands-on use of leading edge tools. For example, in the data management lesson, the students were asked to evaluate a data plan as if they were serving on a review panel. Tools used as part of the other elements included Github, R, and Python.

Science communication, another element addressed early in the three-week process with a similar goal to sensitize participants to some of the pitfalls typically encountered by scientists as they seek to talk about their research. A group exercise to develop a message box (See http://www.scribd.com/doc/139351833/The-COMPASS-Message-Box) related to their research was used as a tool to focus their ideas. The message box concept is an approach to facilitating science communication taught by Compass.[6] The message box was returned to throughout the three weeks. Not only was it helpful in terms of facilitating communication of the research, but also the message box was important as a means of developing and fine tuning their research in the context of the group projects during the course of the three weeks.

The day-to-day operation of the training institute relied on core staff, instructors, onsite technical consultants, and related support staff. Core institute staff and instructors were responsible for the overall intellectual organization and actual institute content. Onsite technical consultants provided real-time, in-person help during the topical exercises and during the group project time. Other staff provided support to ensure that the AV and networking functioned appropriately, and that all other aspects of the institute ran smoothly.

Results

Practical Lessons

A starting premise for developing the joint institute was to experiment with conducting the training as a concurrent bi-coastal activity. In the process of developing the 2014 OSS training institute, the two teams worked to overcome their biggest challenge, the time differences between the East and West Coast time zones. The teams organized the daily schedule in such a way as to accommodate this difference with the East Coast portion working on their group projects in the morning before the West Coast team joined in, then the West Coast team did their group project work at the end of the day after the East Coast participants adjourned. (See Figure 1) Ironically, another aspect that the organizers confronted was software compatibility and interoperability. Recognizing the tradeoff between participant familiarity with their own laptops and operating systems and the learning curve associated with a course-specific device, operating system and software stack presented via a loaner, the institute organizers also opted for a 'bring your own device' strategy. While this approach had some advantages, the disadvantages became evident, as not all packages were equally functional on all platforms. In future versions of the training institute, the organizers

[6] http://www.compassonline.org

may create a preconfigured software stack for use by the participants accessible via logging into a hosted virtual environment. Another area of challenge for the joint institute was relying on web video conferencing technology. The teams utilized three screens, one screen for presenter material, one screen featured a video feed of the presenter, and one screen showed a video feed of the remote set of participants. A teleconference line was the standby backup in the event the web-based videoconferencing failed. Other important elements to be noted also include the need to provide significant amounts of hands-on support as the participants work through the exercises presented by the instructors.

Assessment and results

In order to increase the immediate and long-term benefit of the training for the participants, three feedback elements are included as part of the institute approach. The first is a pre-course participant survey, the second is a daily debriefing session with participants, and the third is a set of assessment surveys one mid-course and one administered post-institute. The pre-course participant survey is intended to obtain information about skill levels, technical interests and research interests. The goal is to help shape the content for the institute to meet participant needs. Daily debriefs with participants offer opportunities to identify areas that may need more in-depth treatment or additional topics to address in follow-on sessions which gives the institute an overall feel of a highly participatory advanced seminar. Finally, the mid-point and post-institute assessment are used to guide to make improvements in the second half of the course and to guide future implementations of the institute as well as to assess scalability of the approach.

Based on preliminary assessment results from the first institute run by NCEAS during 2013, students reported significant increases in their confidence with respect to many aspects of collaboration research involving enhanced computational approaches - from sociological to confronting the technical challenges of data sharing. They also reported dramatic increases in comfort in working at the command line, archiving data, using version control software, and creating a relational database, and there was a significant but less dramatic increase in their comfort with integrating heterogeneous data. There was no change in their attitudes toward open science, but this is partly due to a high starting average. In addition to self-assessment, we asked a few narrative questions that were then graded, for a more objective measure of what they had learned. Scores improved significantly for their descriptions of organizing information in a collaboration, using "version control" in their programming, and creating a relational database without proprietary software; they also had a small (starting scores were high) but significant increase in their scores for working with spatial data without proprietary software. There was not a significant change in scores for creating metadata standards (and average scores were low).

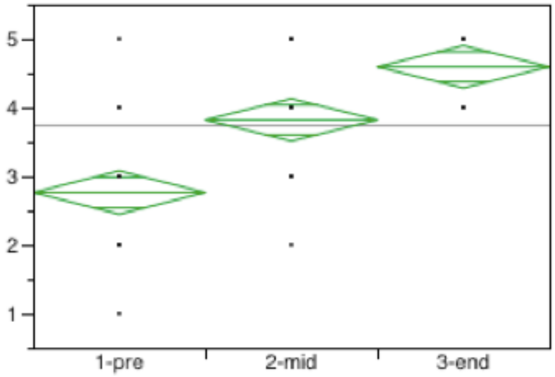


Figure 2: Representative Question Regarding Perceived Ability to Share Data in a Large Collaboration (5 represents increased ability)

In comparison to the 2013 evaluation, the approach used during the 2014 OSS Institute asked participants to self-evaluate their proficiencies in various areas at the beginning of the training and then asked participants to do another self-evaluation at the end of the course to report their perceptions of changes in professional development resulting from participation in the course. Each evaluation addressed the same topics. The topics were grouped in a way that corresponded to the main teaching themes of the institute including Analytical (Statistical) Approaches, Statistical

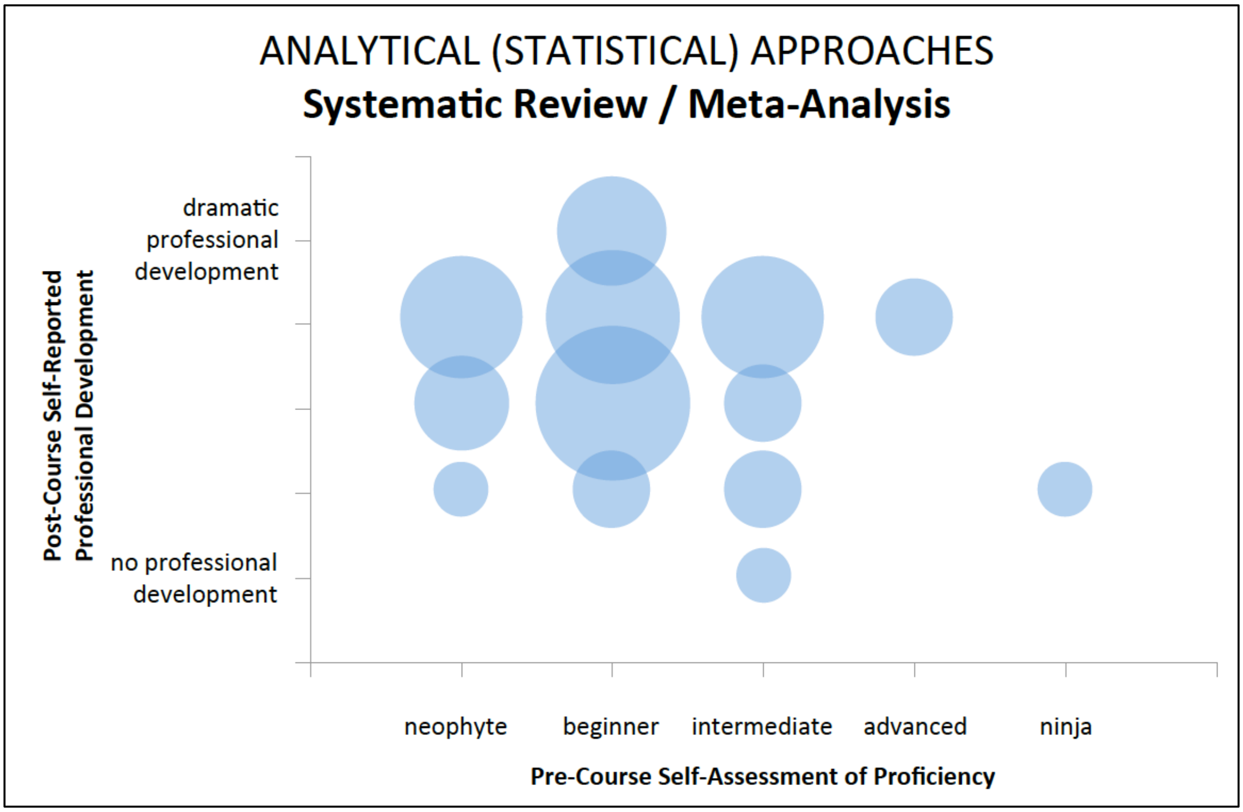
Design and Analysis Tools, Command Line Computing, Software Design and Architecture, Data Visualization, Data Management, Scientific Synthesis, Collaborative Research, and Professional Networking Skills. In the assessment, these categories were subdivided into individual topics such as Systematic Review and Meta-Analysis under Analytical Approaches, Documenting Data for Interpretation and Use by Other Researchers under Data Management, Use of Specialized Online Collaborative Tools under Collaborative Research, and so on. We have included two charts in this paper representative of this analysis. See Figure 4. Overall, the results were generally positive indicating that students gained relevant knowledge and experience from participation in OSS 2014. Though not all categories showed equally strong effects. However, this likely reflects the challenge of covering all topics in equal depth. A full description of these results is the subject of a separate paper.

Figure 3: OSS 2014 Digital Badge

Finally, we would also like to note the collaborative process inherent in the OSS approach. As indicated above, the institute utilized ongoing feedback from participants, particularly via a morning debrief session, as a version of continuous process improvement. Participants were able to highlight areas for more work, but they also provided new ideas and insights. For examples, participants put forth the idea of developing a digital badge for OSS 2014. Digital badges are gaining currency as a way to highlight skills and training in online profiles and digital CVs. Participants worked with organizers to develop an OSS 2014 digital badge (Figure 3) that a number of students have added to their online professional profiles.

Conclusions

The goal of this paper is to present a brief overview of an innovative and comprehensive approach to addressing computational training needs in the ecological and environmental sciences in order to further goals related to sustainable software in support of science and open synthesis science. OSS 2014, a joint partnership of two NSF-funded projects - the Institute for Sustainable Earth and Environmental Software (ISEES) and the Water Science Software Institute (WSSI), integrates training related to sustainable science data, collaboration tools, the sociology of collaboration, computational approaches, and sustainable software while at the same time enabling participants to engage in actual collaborative research. Results from direct assessments with the participants in the 2013 and 2014 versions of the institute show the value of the approach. In the future, in addition to refining the approaches deployed in OSS 2014 for the next version of OSS in 2015. We will also seek an opportunity to do additional follow-up with participants to determine the long-term scientific and career impacts of this training. As one participant summarized the experience, “What we learned really ought to be taught to every incoming PhD student in an ecology program, but it just isn't. I feel incredibly lucky to have had the opportunity to learn it now!” This underscores the need, the goal, and the positive result for the course.



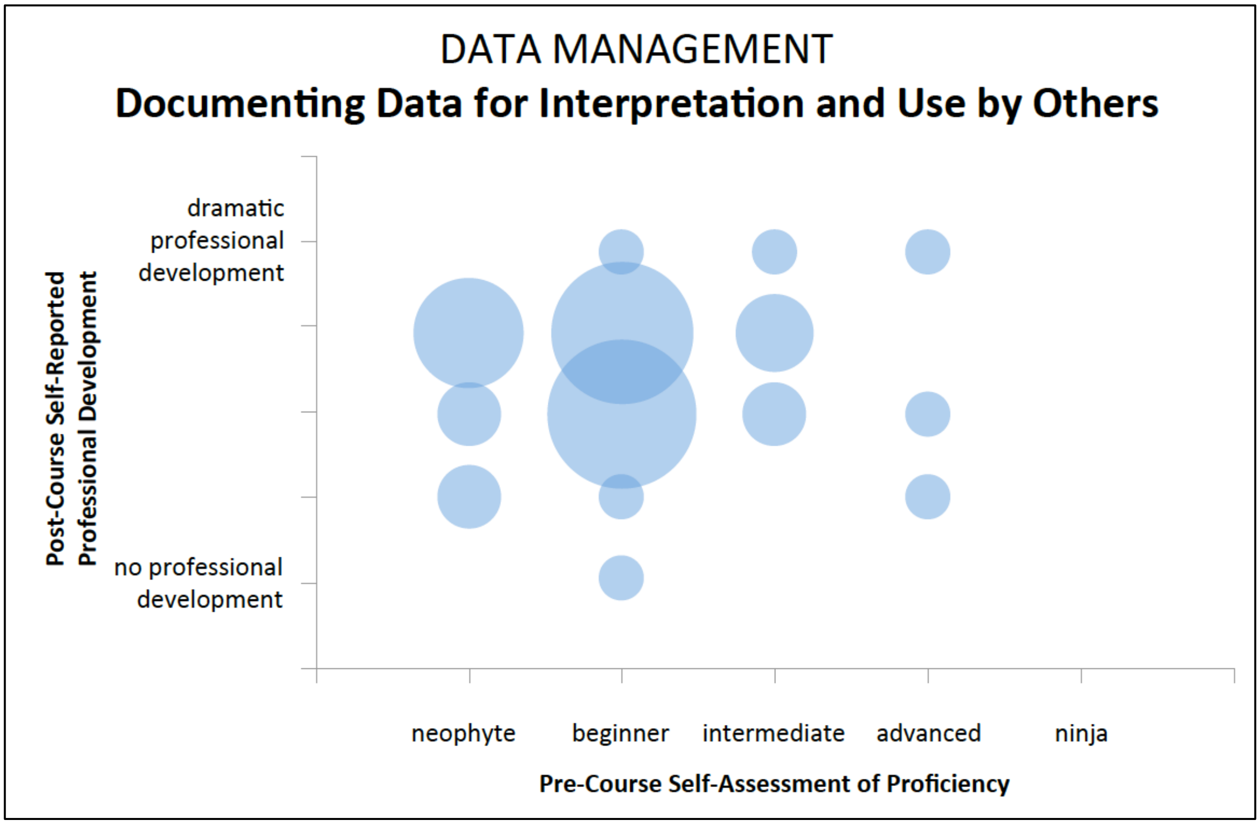


Figure 4: Representative Results from Two 2014 Self-assessment Questions

References

Ahalt, S.; Band, L.; Christopherson, L.; Idaszak, R.; Lenhardt, C.; Minsker, B.; Palmer, M.; Shelley, M.; Tiemann, M.; Zimmerman, A. "Water Science Software Institute: Agile and Open Source Scientific Software Development." *Computing in Science & Engineering*. (IEEE, PP, 2014). DOI: 10.1109/MCSE.2014.5.

Baptista A, Howe B, Freire J, et al. 2008. "Scientific exploration in the era of ocean observatories." *Comp Sci Eng* 10: 53–58.

Hamilton, Michael P.; Graham, Eric A.; Rundel, Philip W.; Allen, Michael F.; Kaiser, William; Hansen, Mark H.; and Estrin, Deborah L. "New Approaches in Embedded Networked Sensing for Terrestrial Ecological Observatories." *Environmental Engineering Science*. March 2007, 24(2): 192-204. http://dx.doi.org/10.1089/ees.2006.0045.

Hannay, J.E.; Langtangen, H.P. ; MacLeod, C. ; Pfahl, D. ; Singer, J. ; Wilson, G. "How do scientists develop and use scientific software?" Published in *Workshop on Software Engineering for Computational Science and Engineering*, 2009. SECSE '09. ICSE. http://dx.doi.org/10.1109/SECSE.2009.5069155.

Hernandez, Rebecca R., Matthew S. Mayernik, Michelle L. Murphy-Mariscal, and Michael F. Allen. "Advanced Technologies and Data Management Practices in Environmental Science: Lessons from Academia." *BioScience* 62, no. 12 (December 1, 2012): 1067–76. http://dx.doi.org/10.1525/bio.2012.62.12.8.

Katz, D.S., Choi, S.T., Lapp, H, Maheshwari, K, Löffler, F, Turk, M, Hanwell, M.D., Wilkins-Diehr, N, Hetherington, J, Howison, J, Swenson, S, Allen, G.D., Elster, A.C., Berriman, B and Venters, C 2014. "Summary of the First Workshop on Sustainable Software for Science: Practice and Experiences (WSSSPE1)." *Journal of Open Research Software* 2(1):e6, DOI: http://dx.doi.org/10.5334/jors.an

Keller, Michael; Schimel, David S.; Hargrove, William W.;and Hoffman, Forrest M. 2008. "A continental strategy for the National Ecological Observatory Network." *Frontiers in Ecology and the Environment* **6**: 282–284. http://dx.doi.org/10.1890/1540-9295(2008)6[282:ACSFTN]2.0.CO;2

Lenhardt, W.C., Ahalt, S, Blanton, B, Christopherson, L and Idaszak, R 2014. Data Management Lifecycle and Software Lifecycle Management in the Context of Conducting Science. *Journal of Open Research Software* 2(1):e15, DOI: http://dx.doi.org/10.5334/jors.ax.

Merali, Zeeya. "Computational science: ...Error…why scientific programming does not compute." Published online 13 October 2010. *Nature* **467**, 775-777 (2010). doi:10.1038/467775a

Newton, Adrian C.; Hill, Ross A.; Echeverría, Cristian; Golicher, Duncan; Benayas, José M. Rey; Cayuela, Luis; Hinsley, Shelley A. "Remote sensing and the future of landscape ecology." *Progress in Physical Geography* August 2009 33: 528-546, http://dx.doi.org/10.1177/0309133309346882.

Pfeifer, M., Disney, M., Quaife, T. and Marchant, R. (2012), "Terrestrial ecosystems from space: a review of earth observation products for macroecology applications." *Global Ecology and Biogeography*, 21: 603–624. http://dx.doi.org/10.1111/j.1466-8238.2011.00712.x.

Strasser, C.A. and Hampton, S. E. 2012. "The fractured lab notebook: undergraduates and ecological data management training in the United States." *Ecosphere* 3:art116. http://dx.doi.org/10.1890/ES12-00139.1