Title: Data Management and Computational Skills Training for LTER Scientists

Proposers: Dr. Tracy K. Teal[1], Dr. Ethan White[2], Dr. Greg Wilson[3]
[1]Michigan State University, Kellogg Biological Station LTER (tkteal@msu.edu)
[2] Utah State University, Sevilleta LTER (ethan@weecology.org)
[3] Mozilla Foundation/Software Carpentry (gvwilson@third-bit.com)

Background and Goals

The LTER network is an essential resource for the study of ecological systems. This network of sites has collected a broad array of data at a combination of spatial and temporal scales that is largely unprecedented in ecology. With multi-decadal data from dozens of sites across North America there is enormous potential for long-term analysis, cross-site synthesis and the integration of diverse data types. LTER datasets currently include everything from gas flux data to soil chemistry, GIS and genomic data. However, this volume and heterogeneity of data presents challenges for data managers and scientists interested in managing and utilizing this data.

As science has becomes increasingly computational, the education gap between the skills that most scientists *have* and the skills that they *need* continues to widen. The hidden costs of this are painful: tasks that should take minutes wind up taking hours, insights are missed, and collaboration is impeded. The root cause is a lack of the basic skills that allow scientists to create and customize software and effectively utilize computational approaches more generally. Even parsing data so that it is in an appropriate format for analysis can be a challenge. Moreover, there is also a general lack of *trainers* and *material* in computational science, making it difficult to overcome this limitation. This is a set of challenges that was identified in several sessions at the LTER All Scientists Meeting in September, 2012.

Education in two areas is necessary to assist data managers and scientists in data management, acquisition, analysis and synthesis. The first is basic computational practice, ranging from effective computer use to advanced topics such as reusing and creating code and highly concurrent programming. The second is data use, archiving and reuse, together with effective metadata generation, storage, and analysis. The LTER network and its data managers have strong backgrounds in data archiving and provision, and so our goal is to help LTER scientists be more productive and make more effective use of LTER datasets by teaching them basic computing skills and a general approach to working in data-driven science.

Activities

We propose a training program to (1) run a number of focused workshops to teach basic computing skills and computational science techniques and train others in delivery; and (2) provide open source self-paced on-line instruction that can cover topics outside of the courses or be self-guided to facilitate self-learning by scientists across a wide range of disciplines.

We will work with the Software Carpentry project (http://software-carpentry.org) to teach short, intensive workshops at multiple LTER sites to train researchers in the core skills needed to be effective in data management, analysis and software practices. Software Carpentry has been working to address these problems since 1998. Over 14 years, it has demonstrated that a modest investment in training can increase scientists' productivity significantly, while making their technology-based work more reliable and shareable. Two independent assessments done in the spring of 2012 reported significant increases in participants' productivity. The key is to focus on fundamental skills such as version control, automated testing, data management, task automation, and program design. These are approaches that empower scientists to solve today's problems efficiently and tackle new ones tomorrow.

The Software Carpentry group also provides many beginning- and mid-level static and video tutorials suitable for scientists who want to become more proficient in scripting, command-line, and other basic skills. All of the material is under a Creative Commons license and is hence reusable. The material underpins the twenty-one workshops that Dr. Wilson and others have run over the last year with seed funding from the Sloan Foundation. The material has been used to teach scientists across the physical, biological, and social sciences. Dr. Ethan White and Dr. Tracy Teal, co-PIs on this proposal, have been involved in the development of materials and in teaching short courses. Both have experience in ecological sciences and bioinformatics. Dr. White studies general patterns in ecological systems at continental to global scales,. while Dr. Teal's focus has been on microbial ecological and metagenomic analyses. They will be involved in tailoring Software Carpentry material to the LTER community. Additionally, there is support from others in the LTER community not listed as PIs on this proposal, but who have had collaborators or students involved in Software Carpentry courses.

Our workshops will rely primarily on face-to-face workshops, because while online education has made significant strides in the past decade, it is still markedly less effective in transferring understanding to novices than direct instruction. These workshops will be hands-on, interleaving short tutorials with live coding sessions, and will build on the material and experience of the Software Carpentry project, the Next-Generation Sequencing course run by Prof. Titus Brown (Michigan State University), and Dr. White's Programming and Database Management for Biologists courses.

Our basic curriculum will introduce:

- The Unix shell. This will show participants a minimal set of useful commands, but the real aim is to introduce them to the ways in which interactive interpreters can accelerate their scientific research.
- Proper program design using either Python or R (the audience's choice).
- Version control for file sharing, collaboration, and reproducibility.
- Quality assurance. This starts with the idea of unit testing, but extends to designing testable code and using test-driven development.
- One additional topic chosen to meet the specific needs of workshop attendees, such as an introduction to databases and SQL.

Participants and Sites

We will run six two-day workshops at LTER member institutions in 2013, working with the LTER Network Office to identify appropriate sites. We will train 40 participants per workshop, and will involve more experienced practitioners in the workshop area as helpers (in part to "train the trainers"). We will strive to include data and information managers as well as graduate students and postdoctoral researchers engaged in research at the LTERs. If demand for the courses exceeds available space, we will try to balance the levels of expertise and project diversity. Having multiple workshops will allow us to maximize the number of institutions and LTER sites represented while keeping travel costs for participants down. We will encourage participants to take part in groups, e.g., at least 4-5 people from a site, so that they are less inhibited about asking questions, and can support each other afterward. Involving participants from multiple projects and expertise levels can also help foster cross-site and interdisciplinary collaborations. Course timing is flexible and can be managed to meet the needs of the LTER Network Office and particular LTER sites.

Expected Outcomes

The ultimate measure of success is whether scientists can do more research in less time and tackle problems they could not have tackled before. As both are difficult to measure, especially in the short term, we will use pre- and post-workshop questionnaires and interviews to gauge what impact the training has had (both qualitative and quantitative).

Budget

All instructors are volunteers, so the costs will be limited to travel expenses for instructors and on-site workshop costs (catering, equipment, etc).

The proposed budget is for six workshops. If demand for the courses were high, additional workshops could be taught.

Expenses for one workshop

Cost	***Type***	***Description***
$3200	Travel/accommodation for two instructors	$1600 per instructor with two instructors per course
$1000	Workshop costs	Equipment, food for lunches, etc

Evaluation expenses (one time only)

Cost	***Type***	***Description***
$5000	Salary	100 hours of a graduate student's time to conduct the evaluations. Part of a broader evaluation initiative by Software Carpentry.

Total		
$30,200	Six workshops and evaluation expense	