IDCC15 | *Practice Paper*

Data Carpentry: workshops to increase data literacy for researchers

Tracy K. Teal
Michigan State University, East Lansing, MI, USA

Karen A. Cranston
National Evolutionary Synthesis Center (NESCent), Durham, NC, USA

Hilmar Lapp
National Evolutionary Synthesis Center (NESCent), Durham, NC, USA

Ethan White
Utah State University, Logan, UT, USA

Greg Wilson
Software Carpentry Foundation, Toronto, Canada

Aleksandra Pawlik
University of Manchester, United Kingdom

Abstract

In many domains of science the rapid generation of large amounts of data is fundamentally changing how research is done. The deluge of data presents great opportunities, but also many challenges in managing, analyzing and sharing data. Good training resources for researchers looking to develop skills that will enable them to be more effective and productive researchers are scarce. To address this need we have developed an introductory fully hands-on workshop, Data Carpentry, designed to teach basic concepts, skills, and tools for working more effectively with data.

Using the highly successful Software Carpentry workshops as a model, we developed Data Carpentry as a two-day workshop for which we used the existing novice training materials from Software Carpentry, modified for our own purposes and we developed new lessons. The materials are designed to facilitate learning by researchers with little to no prior knowledge of programming, shell scripting, and command line tools.

Many organizations and groups have been working to develop a Data Carpentry

Draft from 13th October 2014

Correspondence should be addressed to Aleksandra Pawlik, Room 1.17 Kilburn Building, Oxford Road, University of Manchester, M13 9PL, Manchester, United Kingdom. Email: aleksandra.pawlik@manchester.ac.uk

The 10th International Digital Curation Conference takes place on 9–12 February 2014 in London. URL: http://www.dcc.ac.uk/events/idcc15/

course, including ELIXIR-UK[1], ANDS[2], and BIO CollabIT[3]) due to the shared need for better training researchers in advanced data management, analysis, and computational literacy. At a BIO CollabIT meeting in September 2013, informatics staff from several of the represented interdisicplinary science centers developed the cornerstones of a data and computational literacy workshop based on the Software Carpentry model.

To attain this objective, we identified the following teaching subjects.

- How to use spreadsheet programs (such as Excel) more effectively, and the limitations of such programs.
- Getting data out of Excel and into more powerful tools – using R or Python.
- Using databases, including managing and querying data in SQL.
- Workflows and automating repetitive tasks, in particular using the command line shell and shell scripts.

In addition to the above subjects, the following skills emerged as particularly important to impart from our discussions about designing the course:

- Preparing data for analysis.
- Using data and computational resources, in particular publicly available ones such as Amazon Web Services or iPlant Atmosphere.
- Conducting data and computation-heavy research more reproducibly and openly.

The first Data Carpentry workshop took place at NESCent May 8-9, 2014[4]. The course provided room for 30 learners, and it was full within 3 hours of opening registration. An additional 64 people registered for the wait list, and anecdotal evidence suggests that there were many people who were interested but did not register for the wait list once they saw that the course was full. Post-assessment ratings gave the course an average rating of 8.25 out of 10. The experience of teaching the material and feedback from learners also gave rise to ways of refining both the topics taught and the order in which they are taught, as discussed in a post-workshop write-up[5].

The second workshop took place at BEACON in July 24-25, 2014[6]. The third workshop took place at iDigBio in 29-30 September, 2014[7]. We were able to use the new materials covering the topics such as: working with spreadsheets, SQL, OpenRefine and R. By using the same dataset trhoughout all modules we were able to provide a The first Data Carpentry workshop for ELIXIR-UK is planned for 27-28 November 2014 at the University of Manchester.

Building off the enthusiasm and momentum of Data Carpentry, we want to continue training researchers in good data analysis and management practices and move

[1] http://elixir-uk.org/
[2] Australian National Data Service; http://www.ands.org.au/
[3] An informal consortium of science-supporting IT groups at interdisciplinary centers funded by the NSF BIO directorate; http://www.nescent.org/wg_collabsci/
[4] Course website at http://nescent.github.io/2014-05-08-datacarpentry/
[5] http://software-carpentry.org/blog/2014/05/our-first-data-carpentry-workshop.html
[6] Course website at http://datacarpentry.github.io/2014-07-24-beacon/
[7] Course website at http://datacarpentry.github.io/2014-09-29-iDigBio/

forward with more workshops and organizational support for instructors and materials development to meet these goals.

As the initial design of the workshop resulted from data management training gaps identified as common among the NSF BIO-funded science centers, running more workshops at other centers is the logical next step for Data Carpentry. This will help further polish the materials so that the course meets learners' needs at different centers, in different fields of (biological) science.

Towards that aim there are several longer term goals:

- The ability to host workshops in many different locations, including running them multiple times in the same location
- Materials developed for domains of interest
- Materials in both R and Python
- A streamlined assessment where we can assess learning in a workshop and its effects as researchers progress through their careers
- A forum for continued engagement on Data Carpentry post workshop
- A set of resources where learners could look for more information on particular topics
- More advanced workshops - data visualization, more advanced R or Python for statistics

Several components are already in place or under active development:

- Github repositories for the development and distribution of materials
- Online Software Carpentry tutorials on many topics
- First set of materials adapted from Software Carpentry for the Data Carpentry workshop at NESCent

The components and capabilities yet to be established include the following:

- Personnel for establishing workshop guidelines and structure, materials development and coordinating efforts
- Train-the-Trainers workshops to expand the group of instructors
- The development of assessment materials and mechanisms to assess the impact that Data Carpentry makes

One idea is to operate Data Carpentry similar to a franchise, with a strong Train-the-Trainers component that would enable instructors to run workshops under the Data Carpentry brand at their local institutions, with central logistical support for registration, coordination of material development, and other tasks that benefit substantially from economies of scale.

Acknowledgements

We are grateful for the support of several organizations that contributed personnel time, materials, and travel funds. Specifically, DataONE[8] (NSF #083094); NSF[9](NSF Career award); BEACON[10] (NSF); SESYNC[11] (NSF DBI-1052875), iDigBio[12] (NSF) and iPlant[13] (NSF).

We are also grateful to Software Carpentry and the Mozilla Science Lab for guidance on materials development and administrative support.

[8] http://dataone.org
[9] http://nsf.gov
[10] http://beacon-center.org/
[11] http://sesync.org
[12] http://idigbio.org
[13] http://iplantcollaborative.org