ISSN 2161-3974

JeSLIB 2018; 7(3): e1152
doi:10.7191/jeslib.2018.1152

Journal of eScience Librarianship
putting the pieces together: theory and practice

Full-Length Paper

Integrating Data Science Tools into a Graduate Level Data Management Course

Pete E. Pascuzzi and Megan R. Sapp Nelson

Purdue University, West Lafayette, IN, USA

Abstract

Objective: This paper describes a project to revise an existing research data management (RDM) course to include instruction in computer skills with robust data science tools.

Setting: A Carnegie R1 university.

Brief Description: Graduate student researchers need training in the basic concepts of RDM. However, they generally lack experience with robust data science tools to implement these concepts holistically. Two library instructors fundamentally redesigned an existing research RDM course to include instruction with such tools. The course was divided into lecture and lab sections to facilitate the increased instructional burden. Learning objectives and assessments were designed at a higher order to allow students to demonstrate that they not only understood course concepts but could use their computer skills to implement these concepts.

Results: Twelve students completed the first iteration of the course. Feedback from these students was very positive, and they appreciated the combination of theoretical concepts, computer skills and hands-on activities. Based on student feedback, future iterations of the course will include more "flipped" content including video lectures and interactive computer tutorials to maximize active learning time in both lecture and lab.

Correspondence: Pete Pascuzzi: ppascuzz@purdue.edu

Keywords: data information literacy, flipped instruction, backwards design, active learning, graduate education, STEM education, research data management, data science, R, RStudio, Excel, Unix

Disclosures: The substance of this article is based upon a poster presentation at RDAP Summit 2018. Additional information at end of article.

Introduction

Teaching research data management to graduate students is a trend within information literacy instruction (Carlson et al. 2011; Qin and D'ignazio 2010; Whitmire 2015). A variety of pedagogies have been used, including one shot guest lectures (Westra and Walton 2015), online courses (Johnston and Jeffryes 2014b), traditional courses (Bracke and Fosmire 2015), workshops (Adamick, Reznik-Zellen, and Sheridan 2012), team-based active learning (Clement et al. 2017), and blended learning (Zhang, Goodman, and Xie 2015).

Initial offerings provided instruction on the basic concepts of research data management (Carlson et al. 2011; Qin and D'ignazio 2010). Over time, the educational objectives have become more specific, either by discipline (e.g. agriculture (Carlson and Bracke 2015), engineering (Holles and Schmidt 2018; Johnston and Jeffryes 2014a), meteorology (Frank and Pharo 2016), business (Macy and Coates 2016), climate sciences (Thielen et al. 2017), GIS (Widener and Reese 2016), architecture, history, and social work (Addison, Aaron and Moore, Jennifer 2015), or by skill set (visualization (Konkiel, Marshall, and Polley 2013), personal health data (Johnson 2017), and personal digital archiving (Mannheimer and Banta 2018). Active learning is often an integral part of this instruction to provide students a hands-on experience.

A common limitation of the active learning problems is that students must use data science tools with significant shortcomings because there is no time to train them to use more robust tools. For example, you can teach students to do data validation or visualization with Excel, but it has critical shortcomings in documenting workflows or preventing data corruption. In practice, students may leave the course, workshop or lecture with a list of theoretical data management concepts they need to implement, but they lack the skills to apply these concepts holistically.

This article describes the redesign of an introductory, graduate-level data management course that was developed by librarians at a Carnegie R1 University (Carlson and Stowell Bracke 2015). The original course was offered in collaboration with a department in the life sciences with disciplinary faculty and librarians as instructors. Recent changes at the University included the ability of the library to offer credit courses independent of other departments, and a University-wide data science initiative. These changes warranted a fundamental reimagining of the course rather than simple revision with a major goal of providing students some training with data science tools that they can use to address their data management needs.

Course Development

Previous experiences with the original course informed the course development. The original course covered the fundamental concepts of research data management using an active learning approach. However, given the constraints of the course, the active learning activities typically did not use the tools common to the students' research discipline. For example, file and directory names were developed and critiqued based on the capacity for a person to determine the contents of the file. The only tools that are required for this exercise are a whiteboard. However, file and directory names can contain important metadata that can be parsed computationally and incorporated into data analysis workflows (Noble 2009). This latter concept is difficult to teach unless students can actually manipulate the file and directory names with tools such as Python, R, and UNIX. Thus, there was a clear opportunity to provide some training with computational tools to enhance the active learning activities of the original course.

The liaison experiences and skills of the course developers dictated software and tool selection for the course. These experiences include instruction at the graduate level and research consultations with both graduate students and faculty spanning multiple departments in six colleges in the life sciences and engineering, so the developers are very familiar with the tools utilized by the target students. Another major factor in software and tool selection is that it should be open-source or widely available. Importantly, one developer has extensive experience with computational tools used in the life sciences including Excel, R, and UNIX. Therefore, these tools were chosen for course.

From inception, the course was targeted to graduate students in disciplines that generate quantitative data such as the life sciences and engineering. Course development used a backwards design philosophy (Wiggins and McTighe 2005). Learning objectives were identified for the course as a starting point for creating learning and assessment activities, and include knowledge and skills from the DIL competencies:

- Students will implement current best strategies for their own research project to address the major goals of research data management.
- Students will learn how to efficiently use a computer as their primary data management tool based on knowledge of how files and data are stored on Windows, Mac, and UNIX computer systems.
- Students will devise efficient strategies to collect, store, organize, analyze and share their research data dictated by basic data types, dataset organizations, and best practices for data publication.
- Students will adapt and apply existing metadata schema to their own research data to maximize data publication.
- Students will discover a wide range of public data repositories so that they can identify, acquire, explore, clean, and incorporate published data into their own research project.
- Students will prepare datasets for publication by identifying appropriate repositories, choosing the best format for data sharing and assigning sufficient metadata.
- Students will devise a data management plan that complies with one or more major government funding agencies.

Course development guidelines at the University specify that graduate-level courses must achieve higher-order learning objectives, so all outcomes were designed at Bloom’s taxonomy levels of apply or above. Importantly, the learning outcomes can be assessed with activities that allow students to implement DIL competencies using computer skills. The DIL competencies were mapped to computer skills (Figure 1) to facilitate this goal.

Given that the course must deliver both theoretical concepts and hands-on computer skills, the developers scheduled the course with both lecture and lab periods on different days. This

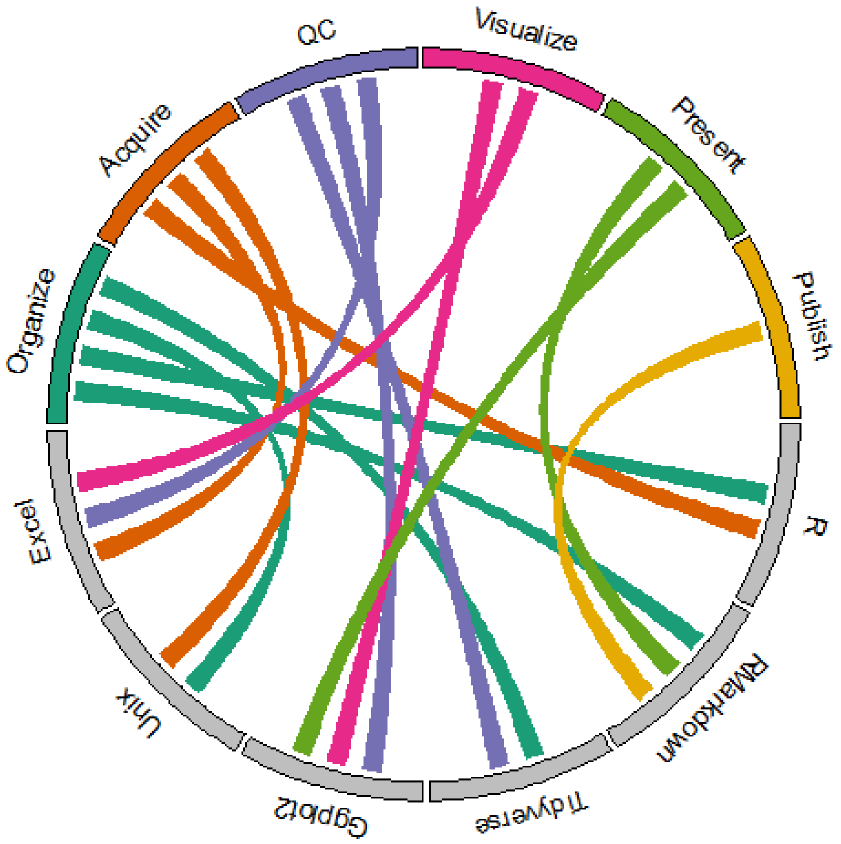

Figure 1: Select Data Information Literacy Competencies Mapped to Computer Skills. The links between data information literacy competencies embedded in the learning objectives to computer skills taught in the lab are explicitly discussed with students to emphasize the strengths and weaknesses of each tool.

allows flexibility for the instruction team, e.g. one instructor can deliver the theoretical material while the other can focus on teaching the computer skills. This also allows students time to internalize concepts before they implement them in the lab.

Lecture

The lecture content provides a theoretical underpinning for the data management activities that the students carry out in the lab. It also gives the students the basic knowledge needed to manage their data outside the class (Table 1).

Each lecture features a hands-on activity that applies the skill that is the focus of the lecture. These activities may apply to data sets, but may also use case studies, discussion, dummy data sets, or educational modules that are publically available such as DataOne (Henkel et al. 2017). The goal of each activity is to stimulate thinking and discussion among the students on the theory behind best data management practices and to evaluate the consequences of bad choices.

Table 1: Topics covered in lectures in chronological order.

What is Data?	Data Acquisition from Public Repositories
The Data Lifecycle	Data Presentation
Storage Media	Data Visualization
Backup Strategies	Data Sharing and Use Agreements
File and Directory Organization	Data Publication
Planning for Data Collection	Big Data
Data Validation	Data Management Plans
Data Exploration and Quality Control	Final Project

Lab

The labs were designed to reinforce concepts from the lecture while providing hands-on instruction with tools commonly used in the disciplines of the target students. A major goal was to help the students use the resources provided by the university to manage their data effectively. Resources included multiple data storage and transfer options, desktop and high-performance computing, and the tools UNIX, R (R Core Team 2017), RStudio (RStudio Team 2016) and Microsoft Excel.

Basic instruction on scripting with R was included and relied heavily on the features of RStudio, especially RMarkdown (Allaire et al. 2018). This file format allowed students to produce documents that incorporated text, figures, file and website links, and executed code that could generated results, e.g. table and figures, that are inserted into the document. Assignments in RMarkdown can be structured to parallel the data lifecycle from data acquisition, cleaning and transformation to analysis, visualization, and sharing (Supplemental Content). Importantly, RMarkdown files are plain text and easy to version control. Currently, RMarkdown supports code written in R, Python, C^{++}, SQL and Stan, and some UNIX commands can be executed with R code. In short, RStudio provides many features that are required by the students to implement their data management strategies.

Many R packages have significant shortcomings when it comes to DIL competencies, and base R syntax for data manipulation is difficult to teach. Therefore, the course used the *tidyverse* collection of R packages (Wickham 2017), including *ggplot2* (Wickham 2009), for several reasons. RStudio provides strong support for these packages in the help menus, making it easier for students to learn. Learning is also facilitated because the *tidyverse* functions are human-literate as opposed to computer-literate, e.g. you can *filter* observations or *select* variables from your data set whereas base R uses a combination of square brackets, dollar signs, and commas to extract data. Importantly, the idea of *tidy* data aligns with data literacy concepts (Wickham 2014), especially issues of data organization, quality assurance and analysis. A shortcoming of *tidyverse* is that it applies to rectangular datasets, i.e. rows and columns of data. This is not a major issues because most of the students work with rectangular data sets, and they can build on their R skills with packages that can handle data formats such as JSON.

Data visualization is a strength of R, and *ggplot2* provides a “grammar of graphics” in which data visualizations are built from multiple layers of data, each of which provides key information to the viewer (Wickham 2010). The idea of layers is not new in graphic design, but the implementation in *ggplot2* is fairly intuitive, easy to teach, and easy to document. Importantly, the data manipulation methods of *tidyverse* integrate smoothly with *ggplot2* data visualization methods.

It should be emphasized that the lab instruction is not intended to teach computer programming, *per se*. Computational concepts such as recursion, nested expressions, and application development are kept to a minimum. Similarly, statistical concepts are not covered in the course, even though R is designed for statistical computing and analysis. The emphasis is on using R and RMarkdown to document processes such as data cleaning, transformation, visualization, and analysis to facilitate research data management.

Some lab instruction with UNIX and Microsoft Excel was also included. At the University, UNIX skills are required to use some computing resources, e.g. logging on, finding software and moving files. Likewise, Excel is thoroughly embedded in most disciplines, so some instruction on best practices with Excel was included, e.g. carefully importing data to avoid corruption, data validation and filtering.

Assessment

Assessments included homework, lab exercises and a data management plan (DMP) with a short presentation. The format of the homework and DMP are typical of DIL courses. The lab exercises are different in that they require students to use computer skills to solve data management problems such as data cleaning or visualization. Students were graded on completing the data management problem, not on the elegance or efficiency of the computer code.

A novel assessment used in the course is the data management notebook written in RMarkdown. The students were instructed to record their notes, specific assignments, R code examples, common computer tasks (e.g. mapping network drives), data management rules (e.g. LOCKSS), and information resources in their notebook throughout the semester. The notebook is the student’s opportunity to integrate the course content into a document that they can readily refer to in the future or share with others. The grading rubric assigned points based on use of good data management, content organization, use of R and RMarkdown, depth and breadth of content, and critical thinking applied to course concepts. Generally, students did quite well with this assignment, and several students produced outstanding notebooks that demonstrated that they were already building on the material covered in the course.

Feedback

Students were surveyed at the end of the course (Supplemental Content). Of 12 students who completed the class, six finished the course survey. The feedback for all aspects of the course was positive. Four were extremely satisfied and two were moderately satisfied with the course, and all respondents were likely to recommend the course to a friend. Students liked the opportunity to apply data management concepts using tools such as R. However, students with less computer experience noted that the infrastructure could be confusing at times. Students also requested more work with case studies in lecture and more time with UNIX and Git version control in lab.

Feedback from faculty in other departments has been informal. There is a general recognition that the training the course provides is important, but getting students in the class is still a challenge. The course was scheduled again for spring 2018, but it was cancelled because of low enrollment. On the success side, for fall 2018, one department required all entering graduate students to enroll in the class. As part of the campus-wide data science initiative at the University, librarians are developing a graduate certificate in data management that will include this course. More formal feedback and discussion of courses will be part of the development process.

Future Developments

Teaching data management is always challenging but adding tools such as Unix and R add complications because students absorb this material at different paces. Computers are intolerant of typographical and syntax errors, so student progress was often hindered by minor errors in coding. This is an unavoidable consequence of computer coding, but it does reinforce the importance of good data management practices such a data cleaning and consistent data entry.

Fortunately, tools such as the *learnr* R package (Borges and Allaire 2018) can be used to develop instructional materials that are amenable to a “flipped” classroom (Brunner and Kim 2016, Brauer, Torfs, and Uijlenhoet 2018). *Learnr* tutorials are small web applications that can be used to teach fundamentals of R programming and data manipulation. Future iterations of the course will have lab instruction supplemented by *learnr* tutorials that students can work through at their own pace. Tutorials have already been piloted on a small scale for other courses.

One limitation of the current course is that engineering students typically do not use R, preferring Python or MATLAB instead. Any course is necessarily limited by the knowledge and skills of the instructors, and the developers of this course have no experience with Python or MATLAB. However, with the proper skills, the lab instruction could be adapted for these tools. Python would be an excellent choice because of its popularity, wealth of packages, and workflow documentation methods such as Jupyter Notebooks.

Conclusions

This course provides students with practical, hands on experience in data management techniques using robust data science tools. Through a lecture/lab format, students get both the theoretical and practical applications that they need and want to manage their research data. Student feedback indicates that the mix of theoretical concepts and practical skills was appreciated. One department has enrolled all new graduate students in the course for fall 2018 and the course is under consideration as actual requirement for that graduate program, showing that departmental support for such courses is possible. Current efforts are directed to scaling-up the course and adapting the content to other disciplines to satisfy a requirement in a nascent data management graduate certificate.

Supplemental Content

Supplemental Files 1 & 2
An online supplement to this article can be found at http://dx.doi.org/10.7191/jeslib.2018.1152 under "Additional Files".

Acknowledgements

The authors would like to acknowledge Jake Carlson and Marianne Stowell Bracke, who developed an earlier iteration of this course in Purdue's College of Agriculture and who left the authors in charge of the earlier course, giving us a chance to iterate. The authors would also like to acknowledge Nastasha Johnson for assistance in developing the LIBR pilot of the course and Noel Diaz and Stephan Miller for developing and supporting the technical infrastructure for the course.

Disclosures

The substance of this article is based upon a poster presentation at RDAP Summit 2018: "Developing an Active Learning Data Course " https://osf.io/b537a.

References

Adamick, Jessica, Rebecca C. Reznik-Zellen, and Matt Sheridan. 2012. "Data Management Training for Graduate Students at a Large Research University." *Journal of eScience Librarianship* 1(3): e1022. http://dx.doi.org/10.7191/jeslib.2012.1022

Addison, Aaron, and Moore, Jennifer. 2015. "Teaching Users to Work Wiht Research Data: Case Studies in Architecture, History and Social Work." *IASSIST Quarterly* 39(4): 39-43. https://doi.org/10.29173/iq905

Allaire, J. J., Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, and Winston Chang. 2018. *Rmarkdown: Dynamic Documents for R*. https://CRAN.R-project.org/package=rmarkdown

Borges, Barbara, and J. J. Allaire. 2018. *Learnr: Interactive Tutorials for R*. https://CRAN.R-project.org/package=learnr

Bracke, Marianne, and Michael Fosmire. 2015. "Teaching Data Information Literacy Skills in a Library Workshop Setting: A Case Study in Agricultural and Biological Engineering." In *Data Information Literacy: Librarians, Data, and the Education of a New Generation of Researchers*, 129-148. Purdue University Press. http://www.jstor.org/stable/j.ctt6wq2vh.11

Carlson, Jake, and Marianne Stowell Bracke. 2015. "Planting the Seeds for Data Literacy: Lessons Learned from a Student-Centered Education Program." *International Journal of Digital Education* 10(1): 95-110. https://doi.org/10.2218/ijdc.v10i1.348

Carlson, Jake, Michael Fosmire, Chris C. Miller, and Megan Sapp Nelson. 2011. “Determining Data Information Literacy Needs: A Study of Students and Research Faculty.” *Portal: Libraries and the Academy* 11(2): 629-657. https://doi.org/10.1353/pla.2011.0022

Clement, Ryan, Amy Blau, Parvaneh Abbaspour, and Eli Gandour-Rood. 2017. “Team-Based Data Management Instruction at Small Liberal Arts Colleges.” *IFLA Journal* 43(1): 105-118. https://doi.org/10.1177/0340035216678239

Frank, Emily P., and Nils Pharo. 2016. “Academic Librarians in Data Information Literacy Instruction: A Case Study in Meteorology.” *College & Research Libraries* 77(4): 536–552. https://doi.org/10.5860/crl.77.4.536

Henkel, Heather, Viv Hutchison, Carly Strasser, Stacy Rebich Hespanha, Kristin Vanderbilt, Lynda Wayne, Stephanie Hampton, et al. 2017. “DataONE Education Modules.” *DataONE*. https://www.dataone.org/education-modules

Holles, Joseph H., and Larry Schmidt. 2018. “Implementing a Graduate Class in Research Data Management for Science and Engineering Students.” In *Design and Implementation of Graduate Education, 2018 ASEE Annual Conference & Exposition*. Salt Lake City, Utah. https://www.asee.org/public/conferences/106/papers/21190/view

Johnson, Matthew Weirick. 2017. “Personal Health Data, Surveillance, & Biopolitics: Toward a Personal Health Data Information Literacy.” *Progressive Librarian* 46(Winter 2017/2018): 150-158. http://www.progressivelibrariansguild.org/PL_Jnl/contents46.shtml

Johnston, Lisa, and Jon Jeffryes. 2014a. “Data Management Skills Needed by Structural Engineering Students: Case Study at the University of Minnesota.” *Journal of Professional Issues in Engineering Education and Practice* 140(2): 05013002. https://doi.org/10.1061/(ASCE)EI.1943-5541.0000154

———. 2014b. “Steal This Idea: A Library Instructors’ Guide to Educating Students in Data Management Skills.” *College & Research Libraries News* 75(8): 431-434. https://doi.org/10.5860/crln.75.8.9175

Konkiel, Stacy, Brianna Marshall, and David Edward Polley. 2013. “Integrating Data Management Literacies with Data Visualization Instruction: A One-Shot Workshop.” Presented at the *Data Information Literacy Symposium*, West Lafayette, Indiana, September 22. http://hdl.handle.net/2022/16814

Macy, Katharine V., and Heather L. Coates. 2016. “Data Information Literacy Instruction in Business and Public Health: Comparative Case Studies.” *IFLA Journal* 42(4): 313-327. https://doi.org/10.1177/0340035216673382

Mannheimer, Sara, and Ryer Banta. 2018. “Personal Digital Archiving as a Bridge to Research Data Management: Theoretical and Practical Approaches to Teaching Research Data Management Skills for Undergraduates.” In *The Complete Guide to Personal Digital Archiving*. ALA Editions/Neal-Schuman. https://scholarworks.montana.edu/xmlui/handle/1/12678

Noble, William Stafford. 2009. “A Quick Guide to Organizing Computational Biology Projects.” *PLoS Computational Biology* 5(7): e1000424. https://doi.org/10.1371/journal.pcbi.1000424

Qin, Jian, and John D’ignazio. 2010. “The Central Role of Metadata in a Science Data Literacy Course.” *Journal of Library Metadata* 10(2-3): 188-204. https://doi.org/10.1080/19386389.2010.506379

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org

RStudio Team. 2016. *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, Inc. http://www.rstudio.com

Thielen, Joanna, Sara M. Samuel, Jake Carlson, and Mark Moldwin. 2017. “Developing and Teaching a Two-Credit Data Management Course for Graduate Students in Climate and Space Sciences.” *Issues in Science and Technology Librarianship* 86(Spring 2017). http://dx.doi.org/10.5062/F42Z13HQ

Westra, Brian, and Dean Walton. 2015. “Teaching Ecology Data Information Literacy Skills to Graduate Students: A Discussion-Based Approach.” In *Data Information Literacy: Librarians, Data and the Education of a New Generation of Researchers*, edited by Jake Carlson and Lisa Johnston. Purdue University Press. http://www.thepress.purdue.edu/titles/format/9781612493527

Whitmire, Amanda L. 2015. “Implementing a Graduate-Level Research Data Management Course: Approach, Outcomes, and Lessons Learned.” *Journal of Librarianship and Scholarly Communication* 3(2): p.eP1246. https://doi.org/10.7710/2162-3309.1246

Wickham, Hadley. 2009. Ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. http://ggplot2.org

———. 2010. “A Layered Grammar of Graphics.” *Journal of Computational and Graphical Statistics* 19(1): 3-28. https://doi.org/10.1198/jcgs.2009.07098

———. 2014. “Tidy Data.” *Journal of Statistical Software* 59(10): 1-23. https://doi.org/10.18637/jss.v059.i10

———. 2017. *Tidyverse: Easily Install and Load the “Tidyverse.”* https://CRAN.R-project.org/package=tidyverse

Widener, Jeffrey M., and Jacquelyn Slater Reese. 2016. “Mapping an American College Town: Integrating Archival Resources and Research in an Introductory GIS Course.” *Journal of Map & Geography Libraries* 12(3): 238-257. https://doi.org/10.1080/15420353.2016.1195783

Wiggins, Grant P., Jay McTighe. 2005. *Understanding by Design*. Expanded 2nd edition. Alexandria, VA: Association for Supervision and Curriculum Development. http://www.ascd.org/Publications/Books/Overview/Understanding-by-Design-Expanded-2nd-Edition.aspx

Zhang, Qinqin, Maren Goodman, and Shiyi Xie. 2015. “Integrating Library Instruction into the Course Management System for a First-Year Engineering Class: An Evidence-Based Study Measuring the Effectiveness of Blended Learning on Students’ Information Literacy Levels.” *College & Research Libraries* 76(7): 934-958. https://doi.org/10.5860/crl.76.7.934