The Impact of Targeted Data Management Training for Field Research Projects - A Case Study

Jonathan L. Petters, George C. Brooks, Jennifer A. Smith, Carola A. Haas

Jonathan Petters: Data Services, University Libraries, Virginia Tech, Blacksburg, VA 24061 USA
Corresponding author - jpetters@vt.edu

George Brooks and Carola Haas: Department of Fish and Wildlife Conservation, Virginia Tech, Blacksburg, VA 24061 USA

Jennifer Smith: Department of Environmental Science & Ecology, The University of Texas at San Antonio, San Antonio, TX 78249 USA

Abstract

We present a joint effort at Virginia Tech between a research group in the Department of Fish and Wildlife Conservation and Data Services in the University Libraries to improve data management for long-term ecological field research projects in the Florida Panhandle. Consultative research data management support from Data Services in the University Libraries played an integral role in development of the training curriculum. Emphasizing the importance of data quality to the field workers at the beginning of this training curriculum was a vital part of its success. Also critical for success was the research group's investment of time and effort to work with field workers and improve data management systems. We compare this case study to three others in the literature to compare and contrast data management processes and procedures. This case study serves as one example of how targeted training and efforts in data and project management for a research project can lead to substantial improvements in research data quality.

Keywords

Ecology, Wildlife Conservation, Libraries, Data Management, Field Research

Introduction

The management of research data in its 'long tail' (Heidorn 2008), where data are collected, analyzed and archived by small research groups, continues to challenge researchers and curators. While research in many disciplines has become increasingly data-intensive, allocation of resources for data management has not always kept pace (Brase et al. 2014) and data management training for scientists is lacking (Tenopir et al. 2016). Yet, research funding agencies and publishers have increasingly emphasized data management and sharing. In the United States this emphasis increased with the National Science Foundation's (NSF) 2011 requirement that data management plans be included with research funding proposals (NSF Proposal Guide 2011).

Since NSF's data management requirement was established, research libraries have stepped up to provide research data management services within academic institutions (Fearon et al. 2013). These services typically provide researchers with data management planning support, data management training opportunities, and some also provide curation services by which researchers can share their data (Tenopir et al. 2014). Some of these research data management services are geared towards domain expertise or have domain experts within them (Wittenberg, Sackmann & Jaffe 2018; Teperek et al. 2018), and thus can provide more targeted data management training and support for researchers. For example, at Virginia Tech the research data management service is staffed with PhDs from engineering, social sciences, biological sciences and geosciences (Ogier et al. 2018). While these personnel do not cover all disciplines and subdisciplines of research across the university, it does allow the service unit to provide a level of tailored training for some researchers at our institution.

The first author (Jonathan Petters of Data Services in the University Libraries) helped develop a tailored training for a research group in the Department of Fish and Wildlife Conservation at Virginia Tech. The goal was to build a data management training curriculum for field workers on long-term ecological field research projects in the Florida Panhandle. Here, we showcase this training curriculum and the subsequent efforts of the research group (Carola Haas, George Brooks, and Jennifer Smith, referred to as 'the co-authors' throughout) as a case study in building data management capacity. This case study serves as one example of

how targeted training and efforts in data and project management for a research project can lead to substantial improvements in research data quality.

Case Study

In June 2017, Haas contacted Data Services and outlined several data management issues they wanted to address with respect to their long-term ecological field research projects in the Florida Panhandle. The research team involves faculty, graduate students, technical staff, and undergraduate students, permanently based at Virginia Tech, and field crew leaders and field technicians, permanently based on site in Florida. The work includes population assessments of rare and declining amphibians and reptiles, including the federally endangered reticulated flatwoods salamander (*Ambystoma bishopi*) and the state threatened gopher tortoise (*Gopherus polyphemus*). Both species are at risk of illegal collection for the pet trade, and thus the data, particularly the site locality information, are highly sensitive in nature. Therefore, the databases were not amenable for storage on cloud services such as Google Drive or Dropbox, and instead have been relayed via portable storage devices or over lengthy email chains for several years. Haas's initial e-mail elaborated these issues:

> "I have a large long-term research project that takes place at a remote research site…Over the years, we've had different techs and graduate students creating or adding to databases…now we're in a chaotic situation. Every database I open is full of typos, [field workers] are entering the same data in different places…I've tried to work with them on version control but there are still regular problems with multiple copies of databases going around and one tech is entering into one and another is using a different one."

Petters responded to this initial e-mail and met the co-authors soon after. Their discussion revealed concerns that this "chaotic" data management situation could be compromising the quality of field data used in analysis, and that it could have negative downstream effects on the research. This first meeting resulted in several potential approaches to addressing these concerns. These approaches were:

> a) changing data management applications and hardware,
> b) providing technical training on data management and data management applications,
> c) developing clear written policies and procedures for field data management, and
> d) explaining to the field workers why data management is important to further research goals (i.e. motivating their intrinsic interest in data management).

Over the course of a few meetings it became clear that a primary issue centered around a disconnect between how the field workers collected and managed the field data and how and the research group at Virginia Tech were using the data. Focusing on approaches c) and d) above, the co-authors agreed that the field workers (both seasonal and more permanent) would benefit most from a formal research data management training curriculum prior to November, when the field workers ramp up flatwoods salamander data collection.

Petters developed a one-and-a-half day customized data management training curriculum with the input of the co-authors, and in September 2017 they presented it to several field workers (Petters et al. 2017). This curriculum incorporated important input from the other co-authors and includes:

- Material to help motivate the importance of research data management (see Figure 1),

- Selected modules and parts of modules from the DataONE (Data Observation Network for Earth) educational modules, and
- A proposed framework for more formal data management principles, roles and responsibilities within the research group.

An important aspect of this presentation and ensuing discussion was not to unilaterally impose external rules on the field workers, but rather to begin a dialogue as to how field data collection and management was currently done and how everyone (field workers and the co-authors) could work together to improve the management of the collected data. Changing organizational practices and procedures, including those in research data management, is greatly eased when there is collective agreement on why change and the effort to make change is important (Gilley, Gilley & McMillan, 2009). While all the DataONE educational modules provide data management content that can benefit researchers and field workers alike, material from the three modules of “Data Entry and Manipulation”, “Data Quality Control and Assurance”, and “Metadata” were selected because the co-authors targeted data collection, quality and documentation as particular pain points(DataONE, 2016b; DataONE, 2016c; DataONE, 2016d). Material from “Why Data Management?” were also used to help motivate the discussion (DataONE, 2016a).

Upon conclusion of the presentation and discussion, and in follow-up via e-mail, both the field workers and co-authors expressed their appreciation for the proposed curriculum. Brooks (via video conference) and the field supervisor Kelly Jones (in-person) subsequently held a data management training session for the field workers in November. Setting aside time for group discussions was key, as crew members were able to articulate to us some of the challenges they faced. The field workers were in support of working with the research group in instituting new protocols for data management and acquisition. The co-authors encouraged the field workers to make data quality a priority in their work (sometimes at the cost of collecting more observations) so that their findings could be shared with the broader community. As Whitlock (2011) states, “The central goal to have in mind when archiving your own data is to ensure that a new user, perhaps someone unknown to you working with the data 20 years later, can correctly interpret the results and derive correct conclusions from the data.” If the data are quality controlled and sufficiently documented for this purpose of long-term archiving, they would also meet the needs of the co-authors in their current research.

While the co-authors agreed on the importance of implementing this training curriculum for the field workers, they also expended effort to improve some of the technical aspects of the data management training workflow ((a) and b) above). Previously the field workers were collecting data in multiple spreadsheets and databases on their field computer, leading to quality issues and confusion. Brooks implemented a process of maintaining the database on the server at Virginia Tech that all the field workers can access through an editable front-end. There were initial stumbling blocks to navigate, as varying levels of permissions to the server had to be created for all, and VPN access provided to the Florida crew. Additionally, several training sessions over Skype were required to 1) explain the functionality, and 2) communicate the utility of a split database design (i.e. a database with separate front-end and back-end components). Following implementation, the split database vastly improved version control, and has received resounding praise as a massive improvement by the technicians that routinely enter data and the researchers that have witnessed the transition. Moving all data to a permanent storage system on the Virginia Tech server yielded additional benefits to data security; the server is backed up hourly, and protected by high-end encryption software.

Once the database had been centralized and permanently housed, systematic proofing of historic data became possible. Basic SQL code was created to search for, and amend, typos (e.g., Temperature = 76°C) and inconsistencies (e.g., *yes* vs *YES* vs *Y*). The research group members also created forms within the split database for data entry, with strict criteria for each data field, to prevent similar errors continuing into the future.

For example, depending on when members of our field workers learned frog identification, they might know the scientific name of the southern leopard frog as *Rana utricularia*, *Rana sphenocephala*, or *Lithobates sphenocephalus*, and the six letter ID code in our database was thus sometimes entered as RANUTR, RANSPH, or LITSPH, making it impossible to easily locate all records of this common species. Availability of just one of these codes on the drop-down menu forced everyone to use the same abbreviation. The co-authors continued to remind field workers that the data they were collecting could be very useful for other researchers and conservation managers decades in the future, but only if these other groups could interpret the information.

The co-authors also worked with the field workers to develop written protocols for data proofing. Originally, Haas assumed that field technicians knew how to proof data. However, some of the field workers thought it meant only glancing over the field data sheet to make sure that a day's worth of data had been entered. They did not realize that the values for each entry should be checked by another observer. Once the co-authors understood that training needed to be provided on data proofing, they could write a protocol including ways to quickly check for major errors, such as sorting on a value or using filters in a spreadsheet to look for unusual or extreme values.

In November 2018 and a year removed from working intensely with the other co-authors, Petters reached out to ask them: "What improvements/changes have occurred with respect to data management for your field research projects?" The co-authors' responses suggest a rousing success story. Both the type and frequency of errors (e.g. typographical, inconsistent data entry) seen in their databases have been drastically reduced, and channels of communication between the research group and field workers have been used to much greater effect. The field workers are now recommending modifications to improve database functionality prior to starting data entry as opposed to after data entry. The co-authors' implementation of new data collection procedures has also led the field workers to be more aware of the need to develop their own protocols for recording and proofing data.

Additionally, the co-authors found this effort to improve data management for these long-term ecological field research projects in the Florida Panhandle worthwhile enough to extend to other research projects and into their educational curriculum. Haas noted that they "are starting a new field project in Virginia and are planning on using the [curriculum] materials for that too, and would definitely recommend it to others." Smith added that through this effort they have "certainly become very aware of data management and best practices and [am] dedicating time in my class next semester about data management" as a professor at The University of Texas at San Antonio.

While data management towards these field research projects has substantially improved in the last year, there are issues that continue to require attention. Metadata entry and filenaming has not yet been standardized, for example. The US Geological Survey (USGS), a potential future funder of these research projects, expects metadata to be kept and made available in the Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata or following the International Standards Organization 191xx series of metadata standards, and could be a goal in the near future for the research group (USGS 2019). The USGS provides a list of several tools for creating such metadata (USGS 2019). Additionally, it is important to acculturate new research group members to these data management practices. A concise data management policies and procedures document could be shared with these new members to facilitate this acculturation.

Discussion

Briefly comparing the present narrative case study to other cases described in the literature does not make for rigorous exploratory or critical incident case study research (e.g. Yin 2009). However, comparison of this case study to a few other cases in the literature (Parsons, Brodzik, & Rutter 2004, Burnette, Williams & Imker 2016, Curdt 2018) is useful in illuminating broader generalities about data management processes and procedures across scientific research.

Parsons, Brodzik, & Rutter (2004) and Curdt (2018) describe large field projects; respectively the Cold Land Processes Field Experiment (CLPX) and the Collaborative Research Centre/Transregio 32 'Patterns in Soil-Vegetation Atmosphere Systems: Monitoring, Modeling and Data Assimilation' (CRC/TR32). Both of these field projects acquired relatively large volumes of data in datasets spanning a wide range of spatial and temporal scales, and these data were to be used by several research teams. Consequently these field projects were resourced with dedicated data management support and training. For example, CLPX supported up to four data management specialists in the field for each of its four intensive observation periods (IOPs), and appointed a 'data wrangler' to oversee transfer of CLPX datasets to respective data providers. CRC/TR32 maintained an online database system (TR32DB) during and after the field project, and provided workshops and tutorials for how CRC/TR32 scientists should interact with the system.

In contrast, Burnette, Williams & Imker (2016) and the present case study describe respective research projects managed by and designed for small research groups (i.e. less than 10 scientists), and did not have the resources for dedicated data management support and training. Thus having researchers within the group who have both data management skills and an internal impetus to focus on data management was vital for improved data management. The co-PIs interviewed in Burnette, Williams & Imker (2016) were motivated by previous data management challenges to start on better footing for this research project. They "explicitly sought out someone with superior attention to detail and organizational skills" to help with data management; a substantial portion of this project manager's time was devoted to data management processes and procedures both prior to and during the research project.

The co-authors of the present study put in concerted time and effort that led to substantial improvement in data management. Petters provided consultative guidance and useful data management training framework for the co-authors and field workers, but without the implementation of this framework and other data management improvements made by Brooks, Smith, and the field-based crew described above we would likely not be able to describe this case study as a success. Also unlike the research project described in Burnette, Williams & Imker (2016), data management improvements in the present case study were made mid-stream (i.e. in the midst of years of data collection that has increased in complexity over time). To facilitate these mid-stream changes, Petters's consultative approach intentionally began with motivating field workers on the importance of field data collection and quality control to underpin robust research. While streamlining the technical aspects of data management is an important step towards improving research data management, technical improvements are not sufficient. Haas notes that addressing the disconnect in understanding of goals and needs between field workers and research group members, and framing the importance of getting everyone on board about why data management was important, was a "huge contribution and a great first approach to staff training". For the field workers, making an observation was an accomplishment and provided its own reward while recording the observation in a particular format and proofreading it seemed tedious. If they found a rare species breeding in a new location or returning to a site that had been unoccupied for years after the crew had conducted habitat restoration, they felt that they had accomplished the major goals. However, for the campus-based research group, those observations were only valuable if they could be published and shared with other managers and researchers. The co-authors could not publish the observations if they could not locate the information or easily summarize it.

Regardless of the scale of the research project, clearly defining data management roles and responsibilities was seen as critical. Petters's framework for these roles and responsibilities was perceived as a useful starting point for more clearly defining who was responsible for what regarding data management. The projects described in Parsons, Brodzik, & Rutter (2004) and Curdt (2018), with dedicated data management staff and data management training for researchers, were able to clearly define roles and responsibilities for data management from the outset of data collection. The research group in Burnette, Williams & Imker (2016) explicitly hired a researcher to oversee data management for the project.

One additional point of comparison and contrast is that the two large field projects (Parsons, Brodzik, & Rutter 2004; Curdt 2018) were designed with an express intention of providing for long-term archiving of and access to datasets collected. In Parsons, Brodzik, & Rutter (2004) CLPX data were transferred to the National Snow and Ice Data Center for these actions. The TR32DB system is the preservation and access point for the CRC/TR32 project described in Curdt (2018). Burnette, Williams & Imker (2016) does not describe archiving and preservation plans. The co-authors of the present study do not intend to make all data collected from their long-term ecological field research projects openly available owing to the sensitivity of endangered species data, but are interested in being able to share some of it publically. Research data consultants in libraries like Petters can help small research groups consider ways to make their data openly accessible when appropriate, and can sometimes provide platforms to this end.

Conclusions

This case study in building data management capacity for a small (less than 10 members) research group at Virginia Tech serves as one example of how targeted training in data and project management can lead to substantial improvements in research data quality for field research projects. Consultative research data management support from Data Services in the University Libraries played an integral role in development of this training. The research group members agreed that emphasizing the importance of data quality to the field workers at the beginning of the training discussion was a vital part of its success. Another critical factor towards these substantial improvements was the co-authors internal motivation to see data quality improve, and to take the time to work with field workers and on data management systems to see these improvements through. The co-authors anticipate continuing to give attention to data management and quality as field workers cycle in and out of their long-term projects.

Acknowledgements

We thank Kelly Jones and Steve Goodman (current and former research associates at Virginia Tech) for bringing the field crew perspective to campus and leading the implementation of the training of the field workers, and thank Brandon Rincon and Vivian Porter for ongoing efforts to improve the data acquisition and management process. We thank Mark Parsons of Rensselaer Polytechnic Institute for his encouragement to publish this case study. We also thank he and Nicholas Caruso (currently in Carola Haas' research group) for comments that led to substantial improvements of this manuscript.

References

- Brase, J., Socha, Y., Callaghan, S., Borgman, C., Uhlir, P., & Caroll, B. 2014 Data Citation: Principles and Practice. In Ray J. (Ed.), *Research Data Management: Practical Strategies for Information Professionals*. West Lafayette, IN: Purdue University Press, pp. 167-186.

- Burnette, M.H., Williams, S.C. and Imker, H.J. 2016 From Plan to Action: Successful Data Management Plan Implementation in a Multidisciplinary Project. *Journal of eScience Librarianship*, *5*(1), p.6. http://dx.doi.org/10.7191/jeslib.2016.1101
- Curdt, C. 2019 Supporting the Interdisciplinary, Long-Term Research Project 'Patterns in Soil-Vegetation-Atmosphere-Systems' by Data Management Services. *Data Science Journal*, *18*(1).
- DataONE 2016a *DataONE Education Module: Why Data Management?* Available at https://www.dataone.org/sites/all/documents/L01_DataManagement.pptx [last accessed 11 September 2017].
- DataONE 2016b *DataONE Education Module: Data Entry and Manipulation.* Available at http://www.dataone.org/sites/all/documents/L04_DataEntryManipulation.pptx [last accessed 13 September 2017].
- DataONE 2016c *DataONE Education Module: Data Quality Control and Assurance.* Available at https://www.dataone.org/sites/all/documents/education-modules/pptx/L05_DataQualityControlAssurance.pptx [last accessed 13 September 2017].
- DataONE 2016d *DataONE Education Module: Metadata.* Available at https://www.dataone.org/sites/all/documents/education-modules/pptx/L07_Metadata.pptx [last accessed 13 September 2017].
- Fearon, D., Gunia, B., Pralle, B.E., Lake, S. and Sallans, A.L. 2013 ARL Spec Kit 334: Research data management services.
- Gilley, A., Gilley, J.W. and McMillan, H.S. 2009 Organizational change: Motivation, communication, and leadership effectiveness. *Performance improvement quarterly*, *21*(4), pp.75-94. https://doi.org/10.1002/piq.20039
- Heidorn, P.B. 2008 Shedding light on the dark data in the long tail of science. *Library trends*, *57*(2), pp.280-299. http://hdl.handle.net/2142/10672
- National Science Foundation 2011 *NSF Grant Proposal Guide, Chapter 11.C.2.j. NSF 11-1 January 2011*. Available at http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_index.jsp [last accessed 7 January 2019].
- Ogier, A.L., Brown, A.M., Petters, J., Hilal, A. and Porter, N. 2018 Enhancing Collaboration Across the Research Ecosystem: Using Libraries as Hubs for Discipline-Specific Data Experts. *Practice and Experience in Advanced Research Computing*. https://doi.org/10.1145/3219104.3219126
- Petters, J.L., Haas, C. A., Brooks, G., & Smith, J. 2017 *Eglin AFB Field Projects Data Management Training Curriculum.* http://hdl.handle.net/10919/89070
- Parsons, M. A., Brodzik, M. J., & Rutter, N. J. 2004 Data management for the Cold Land Processes Experiment: improving hydrological science. *Hydrological processes*, *18*(18), 3637-3653. https://doi.org/10.1002/hyp.5801
- Tenopir, C., Sandusky, R. J., Allard, S., & Birch, B. 2014 Research data management services in academic research libraries and perceptions of librarians. *Library & Information Science Research*, *36*(2), 84-90. https://doi.org/10.1016/j.lisr.2013.11.003
- Tenopir, C., Allard, S., Sinha, P., Pollock, D., Newman, J., Dalton, E., ... & Baird, L. 2016 Data management education from the perspective of science educators. *International Journal of Digital Curation*, *11*(1), 232-251. https://doi.org/10.2218/ijdc.v11i1.389
- Teperek, M., Cruz, M. J., Verbakel, E., Böhmer, J. K., & Dunning, A. 2018 Data Stewardship–addressing disciplinary data management needs. https://doi.org/10.31219/osf.io/5w9pj
- US Geological Survey 2018 *Data Management.* Available at https://www.usgs.gov/products/data-and-tools/data-management/metadata [last accessed 15 March 2019].
- Whitlock, M. C. 2011 Data archiving in ecology and evolution: Best practices. *Trends in Ecology and Evolution* 26(2), 61-65. https://doi.org/10.1016/j.tree.2010.11.006

- Wittenberg, J., Sackmann, A. and Jaffe, R. 2018 Situating Expertise in Practice: Domain-Based Data Management Training for Liaison Librarians. *The Journal of Academic Librarianship*, *44*(3), pp.323-329. https://doi.org/10.1016/j.acalib.2018.04.004
- Yin, R.K. 2009 *Case Study Research: Design and Methods*. 4th edn. Thousand Oaks, CA: Sage Publications.