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

The Data Drill: An Opportunity for Researchers to Practice Accessing and Interpreting Data

Sandi Caldrone, University of Illinois Urbana-Champaign, Urbana, IL, USA, caldron2@illinois.edu iD

Yali Feng, University of Illinois Urbana-Champaign, Urbana, IL, USA iD

Abstract

It is difficult to engage researchers in workshops on data management best practices when there are so many other demands on their time and attention. Even when interest is high, attendance is often low. In response to this challenge, the Research Data Service and the School of Social Work at the University of Illinois Urbana-Champaign partnered to develop a new data management learning activity, the data drill. Like a fire drill, the data drill is a safe way to practice a stressful scenario, in this case, accessing and interpreting a dataset. In this paper, we describe how we designed the data drill, discuss the results of three pilot drills we conducted, and outline our plans to improve and expand upon this activity based on our experiences. Each data drill participant selected a dataset they deemed important to their research but that they were not currently using, and attempted to locate, access, and interpret the data during a virtual meeting with one to two librarian facilitators who helped troubleshoot issues as they arose. This allowed participants to stress-test how well their data is organized and documented and provided facilitators with a window into the researcher-data relationship and a unique opportunity to provide highly individualized support with immediate and long-term benefits.

Citation: Caldrone, Sandi and Yali Feng. 2025. "The Data Drill: An Opportunity for Researchers to Practice Accessing and Interpreting Data." *Journal of eScience Librarianship* 14 (1): e851. https://doi.org/10.7191/jeslib.851.

Introduction

Outreach and education around research data management best practices is challenging. Librarians report difficulties engaging researchers in educational activities about data management, particularly workshops (Bishop et al. 2022; Fear 2015; Southall and Scutt 2017). Although workshops have the potential to reach wide audiences, in our experience RSVPs far outstrip actual attendees, and time spent planning such activities is wasted when only a few researchers show up and a workshop becomes a consultation. We need novel ways to engage researchers who want data management support but are not well served by workshops. Downey et al. (2021) have shown that encouraging researchers to share their own experiences with research can be an effective way to have engaging discussions around data management, but this can be difficult in a workshop setting. Consultations are more suited to such personal discussions, but researchers may be reluctant to schedule consultations unless they have pointed questions.

In response to these challenges, the Research Data Service (RDS) at the University of Illinois Urbana-Champaign piloted a new learning activity, the data drill. Like a fire drill, the data drill is a safe way to practice a stressful scenario, in this case, accessing and interpreting a dataset. In 2021 and 2022, the RDS partnered with the university's School of Social Work (SSW) to conduct data drills as virtual meetings between one participant and one to two facilitators to test how well a dataset is stored, organized, and documented. The key innovation of the data drill is that, unlike most data management learning activities, each participant selected a dataset crucial to their work, a dataset either new to them or that they had not used in some time. During the drill, participants were asked to locate, open, and interpret files from the dataset of their choice. Facilitators observed, asked guiding questions, helped participants interpret the data, and made suggestions for improving dataset organization and documentation. In this paper, we discuss the activity design and the results of the three pilot drills and outline our plans to improve and expand upon this activity based on our experiences.

Activity Design

In 2021, the SSW reached out to the RDS to request an interactive and engaging data management training opportunity for their researchers. In response, RDS staff worked with the SSW administration to design the data drill activity to:

1. Help participants test how well their own research datasets are organized and documented in case they need to revisit them in the future.

2. Raise participants' awareness of the importance of organizing and documenting data.

3. Identify specific ways to improve access to and preservation of a dataset identified by the participants as important to their work.

4. Foster positive working relationships between participants and the library.

To create an opportunity to test dataset documentation and not the researchers' short-term memory, each participant was asked to identify a dataset that was important to their research but that they were not actively using. This could be original data they collected themselves or secondary data acquired from another source. The data could be of any size and format so long as they would be able to access it from their computer during a virtual meeting. To recruit participants, the SSW administration reached out to their faculty, staff, and graduate students via email, and three participants were selected based solely on their interest and availability. Two facilitators, the librarian for the RDS and the subject specialist librarian for the SSW, facilitated the drills and followed up with participants. We specifically asked participants not to refamiliarize themselves with the dataset before the meeting. During the meeting, the participants shared their screen and thought aloud while trying to locate, open and interpret files from the dataset they had selected. Facilitators asked clarifying questions and helped troubleshoot issues as they arose. In closing, the facilitators led an informal wrap-up discussion about lessons learned during the activity.

Results

The facilitators conducted three drills between November 2021 and January 2022 with three participants in different career stages: a research program coordinator, an associate professor, and a doctoral student. Each drill lasted about one hour and was conducted over Zoom to respect pandemic protocols and to facilitate screen sharing. To maintain privacy protections participants only shared data on screen after ensuring it did not contain sensitive information, and no files were provided to the facilitators. Meeting recordings were kept for follow-up and research purposes and stored in a secure Box folder only accessible to the facilitators. All three of the datasets selected by participants were in a tabular format but covered a wide range of topics. They included two original datasets collected by the researchers: one from a long-term home visit program that used a vendor-supplied tool to collect data about multiple family members, and one from a randomized clinical trial with data in three languages. The third participant selected data derived from a publicly available dataset of multi-generational, socio-economic data by household.

Although the datasets varied, participants encountered familiar challenges, such as navigating directories and deciphering file names, differentiating between versions of data files, interpreting metadata, determining if blank cells represented null values or missing data, and bridging inconsistencies in how various team members collected and recorded data. Overcoming these challenges required supplemental information, and in all cases that information had to be pieced together from various sources. This included but was not limited to reviewing documentation from the original project design as well as email records of decisions made as research progressed.

Although there were a variety of hurdles, no challenge was insurmountable. Their personal investment in the data and the resulting relationships built led participants to display levels of grit and creativity that are hard to achieve in learning exercises designed around datasets supplied by an instructor. For example, one participant's tabular data included dozens of columns with long, complex, and cryptic headers. This type of

metadata would be highly discouraging to workshop attendees viewing it for the first time as part of a typical exercise. However, this participant was naturally motivated by her involvement in the research. Although she did not remember their meaning offhand, she quickly copied and transposed the headers into a vertical list in a different worksheet to read them more easily and deciphered them effectively by referencing the survey instruments used for data collection.

Challenges for the facilitators were minor and consistent with other data management consultation activities. Scheduling the drills was somewhat difficult given the various demands on researchers' time, although the support of the SSW administration likely made this easier than it otherwise would have been. Focusing on datasets of real importance to the researchers also probably increased their motivation to make time for the drill. The nature of the activity was more challenging for facilitators since it required providing off-the-cuff feedback with no prior knowledge of the dataset. However, having two facilitators, one with data expertise and one with subject-area expertise, helped significantly. The subject specialist had a better understanding of the meaning of the data and how it connected to research and its field implications in social work, and the data librarian was able to assist with more structural and technical questions, such as identifying artifacts left behind when database data was flattened into spreadsheets.

Afterwards, facilitators contacted each participant with a written summary of the data explored, challenges faced, how those challenges were overcome, and recommendations for helpful tools, practices, and resources. Follow up provided a summary for both the facilitators and the participant, allowed the facilitators time to reflect and provide more thoughtful feedback, and created an opportunity for re-engagement and relationship building. Following up with participants proved effective for relationship building even after a significant period of time had elapsed. We followed up with participants roughly a year after completing the activity, which positioned us well to provide support for new data projects they had started in the intervening months. In future data drills, we plan to follow up shortly after the drills with a written summary and an invitation to schedule further consultations, and then also circle back with participants again approximately one year later to inquire about any new data management needs.

Conclusion

In addition to being highly engaging for both the participants and facilitators, the three pilot data drills provided key insights for improving the activity design and implementation going forward. First, although we used Zoom out of necessity during the pandemic, the screen sharing functionality made it possible for all three attendees to easily see the same screen and as such should be used for future drills even when in-person meetings are possible. Second, focusing on a dataset selected by the participant made for a much more engaging experience than a typical workshop built around data chosen by the instructor. In every case, our hour-long meeting was filled with genuine interest, excitement, and cooperative problem solving. Third, selecting a dataset that participants were not currently using made the consultation more interesting by adding an element of mystery and detective work. More importantly, it allowed us to

stress-test data documentation. Although we recommend that researchers begin drafting documentation during the research process, when working closely with a dataset, it can be difficult to imagine what someone less familiar with the data will need to know. For participants using data they had collected, the drills gave them the opportunity to see datasets with fresh eyes, spot gaps in their documentation, and identify key resources to fill those gaps, saving future time and effort and better preparing those datasets for long-term preservation. Finally, we discovered that the primary educational value of the drills is not the data management concepts that come up during the activity but the activity itself. Originally, we thought these exercises would generate learning materials that could be shared more broadly. We planned to edit the videos down into shareable clips to show real world examples of data management challenges and best practices, but ultimately decided against that approach. While participants faced familiar challenges, each experience was so specific and contextual that clips would not be as helpful to others as we originally thought. The real educational value is in the personal and relational nature of the experience. This approach provides insights not only into the data itself but into the relationship between the researcher and their data.

The main weakness of this individualized approach is that it will not scale as well as workshops that can reach a much broader audience. However, the personalized, puzzling, and collaborative nature of these data drills makes them excellent opportunities for data management coaching, and we intend to build on that going forward by incorporating elements of the Association of College and Research Libraries (ACRL) *Framework for Information Literacy for Higher Education* (2016) into our activity design. The ACRL Framework introduces two key components underlying information literacy education: knowledge practices, which are the skills and abilities learners develop, and dispositions, which are the attitudes and beliefs underlying learners' thoughts and behaviors. Our pilot data drills focused on knowledge practices. In our second phase, we will also incorporate dispositions to help researchers extrapolate from their experience in the data drill to reflect on their data management practices and dispositions in general.

Using the ACRL Framework as a guide, we will develop additional questions to help participants identify points of stress in research data management. With a deeper understanding of participants' sources of stress, we can provide more individualized support based on both their knowledge practices and dispositions. Some points of stress may stem from practices that can be improved with tools, methods, or training that save time and effort. Other stressors may be more deeply rooted in how researchers think about data management, and we may be able to help them find alternative ways to approach or reframe the issue without activating the stressors. In either case, we will encourage honest reflection not by focusing on idealized best practices, but by meeting researchers where they are and approaching the consultation with curiosity but not judgment. By identifying skills and tools researchers are currently using, we can introduce realistic, incremental suggestions to improve documentation and prevent information loss and time waste. For example, instead of insisting researchers stop using email to record important decisions among collaborators, we may recommend an Outlook extension like OneNote to make it easier to translate those emails to dataset documentation. Our focus will not be on what the researcher is doing wrong, but on

how awareness and personalized recommendations can improve their relationship with and experience of their data.

Competing Interests

The authors declare that they have no competing interests.

Data Availability

No data was generated by this project.

References

Association of College & Research Libraries (ACRL). 2016. "Framework for Information Literacy for Higher Education." https://www.ala.org/acrl/standards/ilframework.

Bishop, Bradley Wade, Ashley M. Orehek, Christopher Eaker, and Plato L. Smith. 2022. "Data Services Librarians' Responsibilities and Perspectives on Research Data Management." *Journal of eScience Librarianship* 11 (1): e1226. https://doi.org/10.7191/jeslib.2022.1226.

Downey, Moira, Sophia Lafferty-Hess, Patrick Charbonneau, and Angela Zoss. 2021. "Engaging Researchers in Data Dialogues: Designing Collaborative Programming to Promote Research Data Sharing." *Journal of eScience Librarianship* 10 (2): e1193. https://doi.org/10.7191/jeslib.2021.1193.

Fear, Kathleen. 2015. "Building Outreach on Assessment: Researcher Compliance with Journal Policies for Data Sharing." *Bulletin of the Association for Information Science & Technology* 41 (6): 18-21. https://doi.org/10.1002/bult.2015.1720410609.

Southall, John, and Catherine Scutt. 2017. "Training for Research Data Management at the Bodleian Libraries: National Contexts and Local Implementation for Researchers and Librarians." *New Review of Academic Librarianship* 23 (2/3): 303-322. https://doi.org/10.1080/13614533.2017.1318766.