Practical Academic Librarianship: The International Journal of the SLA Academic Division
5(2):77-91, 2015 http://journals.tdl.org/pal

Data Management for Graduate Students: A Case Study at Oregon State University

Maura Valentino
Oregon State University

Michael Boock
Oregon State University

Abstract

On academic campuses, graduate students produce data and that data is often being lost. This paper describes the efforts at Oregon State University to educate graduate students on the value of their data and of preserving it. Graduate students were interviewed and from this information a successful data management workshop was created and has been updated and presented each quarter in the library. Librarians at other academic institutions may be in position to help graduate students on their campuses realize the value of their data and prevent valuable data from being lost. This case study serves as one example of what libraries can provide.

Keywords: data management; data management workshops; data preservation

Corresponding author: Michael Boock michael.boock@oregonstate.edu

Introduction

Academic libraries have long had a role in acquiring, cataloging, preserving, and making available the scholarship created by faculty and graduate students at their universities in the form of published books, journals, multimedia, theses and dissertations, and grey literature. Aside from aiding faculty and students in discovering and accessing published research, academic libraries have had little to do with the creation and dissemination of scholarship prior to its final publication or release. With the rise of scholarly communication-related services in this century, academic librarians are increasingly involved throughout the research cycle with faculty and students and the research they produce, prior to its publication. For example, academic libraries increasingly provide assistance with copyright retention, dissemination of research in open access repositories, publication in open access journals, and assistance with the active management of research data throughout its lifecycle for access and preservation.

This paper describes the early efforts of Oregon State University Libraries & Press (OSULP) to identify and establish relevant data management services for graduate students. The paper begins with a background section describing OSULP group interviews with research faculty about their data and its management as an effort to determine what value-added research data services OSULP might provide. Meeting participants identified graduate student data associated with theses and dissertations as scholarship that was being lost, and faculty participants agreed that this data would be a logical starting point for data management services at OSULP. The paper proceeds to describe OSULP interviews with graduate students, the creation of a data management workshop informed by those interviews, and the development of a credit-bearing graduate data management course.

In their library guide, Texas A&M University Libraries (2013) defines research data management activities as planning, storage and backup, documentation (metadata), file organization and naming, security and privacy considerations, sharing, and archiving and preservation. For many reasons, faculty and graduate students are increasingly motivated to take a more active role in the management of their research data. Funding agencies such as theNational Institutes of Health Office of Extramural Research (2003) and the National Science Foundation (2011) and university offices such as the Duke University Office of Research Support (2007) and The University of Edinburgh Information Services (2014) often require that data be managed and made publicly accessible. Proper data management can also increase researcher visibility and impact (Collie & Witt, 2011), ensure dataset integrity through proper

Corresponding author: Michael Boock michael.boock@oregonstate.edu

documentation, reduce the risk of data loss, reduce time spent locating data, and increase sharing and reuse (Collie & Witt, 2011, p. 166) in a world where "computational capacity and tools… are giving rise to new modes of conducting research" (Steinhart et al., 2008). In clinical trials, responsible data management techniques have been found to decrease the time between trial and market release (Krishnankutty, Bellary, Kumar, & Moodahadu, 2012).

Librarians' skills in organizing, describing, managing, and preserving collections can be useful to institutions and individuals interested in managing research data to increase the efficiency of the data gathering and storage process and to enhance discovery, preservation, and reuse (Delserone, 2008). Erway (2013), in her call to action for universities to adopt data management standards, agrees that libraries are well-positioned to play a key role in data management initiatives. Librarians have subject area and copyright expertise, and libraries have built digital repository infrastructure, services, and expertise that can be used to provide long-term preservation and access to research data. As impartial and independent campus entities, academic libraries are well-positioned to serve as a central service point for helping faculty and students across the university manage, preserve, and publish their research data.

Faculty Conversations

In the spring of 2010, in an effort to better understand Oregon State University (OSU) faculty needs relating to their research data and as a method for determining what services the library should offer in relation to that data, OSULP invited a select group of faculty from a variety of campus departments to three lunch meetings. The faculty were selected on the basis of subject liaison recommendations. Several library faculty also attended the meetings.

Fifteen attendees from a variety of colleges and academic units (College of Forestry, College of Oceanic and Atmospheric Sciences, College of Science, College of Agricultural Sciences, College of Engineering, Research Office) were asked a series of questions about their data and the value that OSULP could provide in relation to their data. Questions were based on those that Purdue University Libraries ask their faculty as part of their data curation profiling work (Witt & Carlson, 2007).

The OSULP discussions with faculty found that they were interested in having the library explore the development of a variety of data management services. New services developed as a result of these discussions include many of those that have become fairly standard

within academic libraries: data curation services; metadata, data organization, and preservation consultation; identifier creation; and the aforementioned data management workshops. In addition, among the primary outcomes of these meetings, OSULP learned that faculty were interested in having the library work with graduate students on the management of data associated with their theses and dissertations. Faculty suggested this as a way for OSULP to become familiar with some of the data produced on campus, perhaps on a smaller and more workable scale than faculty datasets. Faculty also noted that there is often confusion about data ownership, especially data that is generated or collected as part of national grants and data collected by students for their theses and dissertations (Boock & Chadwell, 2010).

Many of the datasets associated with student thesis and dissertation scholarship are lost forever when the student graduates even though many of the datasets are funded by federal grants written and overseen by faculty as principal investigators. At the very least, university resources are used in the gathering and analysis of the datasets. The loss of this data is a problem confronted by other institutions, including the University of Minnesota Libraries (2007). The University of Minnesota study noted that some professors try to teach students to preserve their data so others can use it, but the process is inconsistent across campus.

Carlson and Stowell-Bracke (2013) posit that graduate students are the future of research in their field and, because of this, it is important to teach them the basics of data management (p. 346). Carlson, Johnston, Westra, and Nichols (2013) found that some faculty members do not feel they have the expertise to advise graduate students relating to data preservation activities. They also feel that graduate students are a natural audience for data management education because they are often asked to provide data management not only for their own data but also for other data being generated in the course of faculty research projects with which they are involved (Carlson et al., 2013, p. 210). Since faculty do not always feel comfortable teaching data management to graduate students and we want to send our graduate students to their next institution with an understanding of data management, we believed that graduate students would benefit from a workshop on this topic.

In "At the watershed: Preparing for research data management and stewardship at the University of Minnesota," Delserone (2008) shared results of a series of interviews and follow-up discussions with faculty about how their library could help with data management. The author discovered a lack of understanding among faculty regarding the value of their data to future researchers and suggested that data management workshops would be a good first step in

offering data management services.

OSULP already had a strong relationship with the University's Graduate School in preserving and disseminating student scholarship in the form of theses and dissertations and the occasional accompanying dataset. Beginning in 2006, the Graduate School established a policy that required deposit of theses and dissertations to the ScholarsArchive@OSU (SA) institutional repository managed by the Center for Digital Scholarship and Services (the Center) within OSULP. In addition, the Center began a digitization project to scan legacy theses and dissertations and make them available in the repository. As of December 2013, the repository contained over 17,500 theses and dissertations produced by OSU students or approximately 76% of those ever produced at the university.

There were other reasons OSULP decided to begin working with graduate students to manage their research data. The newly-formed Center was interested in providing scholarly communication-related instruction to graduate students. OSULP also had a long-held interest in disseminating the scholarship of the university's students. For these reasons, the Center decided to learn more about graduate student data management needs and determined that this would provide an opportunity for the Center to learn about value-added services that could be provided to positively affect the preservation and accessibility of this scholarship. The Center could then apply what was learned about student research data and its management to faculty datasets in the future.

Student Interviews—Methodology

Four librarians in the Center were interested in providing data management-related services. As noted, a workshop or seminar for graduate students on data management seemed appropriate as a first step. But what did graduate students need to know and what did they already know?

One of these librarians invited graduate students from a graduate-level Posing Research Questions class to volunteer to participate in interviews with the team of library faculty. Institutional Review Board (IRB) approval was granted by the University and the team selected five students from the class for interviews.

The team developed an interview protocol (Appendix A) with questions divided into

seven categories:

- General.
- Data Description.
- Data Analysis.
- Metadata.
- Data Dissemination and Ownership.
- Data Preservation and Accessibility.
- Attribution.

The interviews began with general questions about the students' degree programs and the general topic of their research and then moved on to more specific questions about their data. For example, how were the students planning to collect their data, and would existing data also be used? The interviews began with general, open-ended questions to allow the students to explain their research in their own way, for example, "describe the data you are collecting; describe how this data fits in with existing data." More specific questions covering the areas of data analysis, storage, organization, and format, as well as data back-up practices, followed. It was not expected that the graduate students would be conversant with metadata practices or terminology, so the tone of these questions was general rather than specific. For example, instead of asking, "what metadata standards are you using?," we asked, "are you aware of metadata standards in place for describing data related to your topic?"

Questions about data ownership, copyright, and dissemination were included, written as if the graduate student would have detailed knowledge of the ownership of the data. It turned out that many students were working with multiple agencies, and ownership and copyright issues related to their data were not always well-understood. Preservation and accessibility are also key components of data management plans so a last section was added for these questions. Some questions were worded as though the graduate student would know the answer and others were designed to determine the students' level of knowledge of a particular topic. In hindsight, it would have been better to word all the questions in the last two sections as "are you aware?" questions such as "are you aware of copyright issues with regard to this data?"

The team decided to have two library faculty interviewers present at each interview. It was hoped that discussions with students would be detailed but also free form and that the presence of two interviewers would increase the likelihood that appropriate follow-up questions would be asked. It was also agreed that the primary investigator would be present

for all interviews to ensure that the interviews were conducted in a consistent manner. The interviews were recorded, transcribed, and shared with the team. Each investigator took notes in case of recording glitches. The first few open-ended questions, such as "describe the data you are collecting" and "how would you describe its relationship to existing data?" were asked as planned, but each graduate student was so enthused to talk about their research that many of the remaining questions were already answered during the course of the graduate student's description of their research.

Student Interviews—Results

The team was aware that, with only five participants, it would be impossible to learn everything about student research data in its variety, but we were able to develop the curriculum of a data management workshop based on the information gathered. Analysis software was not necessary, as the number of interviews was small and the similarities were instantly apparent. First, the types of data collected by students vary a great deal, and there are differences in how and whether existing data is repurposed and integrated with data collected by the student. This is strongly supported in the literature. In the University of Houston study, Peters and Dryden (2011) identified 13 distinct types of data gathered and used by faculty, and many researchers used more than one type of data in the course of conducting their research. In the University of Minnesota Libraries (2007) report, 11 types of data were identified and many common types of data such as spreadsheets and photographs were not even mentioned. Bicarregui et al. (2013) noted that the types of data gathered and used within a particular discipline varied greatly from researcher to researcher (p. 31).

On this basis, the team decided that the content of the data management workshop would have to be general in order to provide value to participants working with a variety of data types from a variety of disciplines. For example, in the workshop that was developed, the use of non-proprietary file formats is suggested and a variety of data types such as spreadsheets, audio and video files, transcripts, and photographs are discussed. Recommendations of non-proprietary formats for all these types of data are included in the discussion. When file-naming schemes are discussed in the workshop, it is not assumed that any type of information must be included in the file or folder name. For example, rather than suggest that dates must be included in the file name, it is taught that the naming scheme is contextual and informed by

the nature of the data collected. Consistency is stressed rather than a particular naming scheme. Similarly, when data preservation is discussed, many types of data repositories, portals, and methodologies for ensuring preservation are mentioned. What is stressed is that the researcher preserve the data in a repository that matches the data created rather than use a particular repository.

The environment in which the graduate students gather their data and analyze it was also found to vary widely. Some campus laboratories in which students work have strict data management procedures in place. Others do not. As a result, the team decided to promote the data management workshop to lab managers who might benefit from the instruction as well as to graduate students. Carlson and Stowell-Bracke (2013) found a similar result when they studied graduate student data management practices at the Water Quality Field Station at Purdue University; the students, often acting as lab managers, used data management practices they had been taught in other laboratory and research situations and the levels of data management practice varied widely (p. 349).

While the differences in the data types and data collection methods used by students interviewed stood out, there were also a number of similarities. All five students understood the benefits of data management even though their understanding of what good data management entails was not fully formulated. Nor did the students understand the implications of good data management. Participants were not aware of a potential library role in relation to the management of their data and seemed surprised to learn of the assistance the library could provide. A review of the literature reveals that in 2007 the University of Minnesota Libraries conducted an extensive survey of faculty and graduate students regarding their research processes. One of their conclusions was that while scientists used library online resources including journals extensively, they did not consider the library a place to come for help with their data management (University of Minnesota Libraries, 2007). The University of Houston analyzed 10 projects in a variety of disciplines to determine what data management practices were in place. They found that researchers did not think to turn to the library for advice on data management. Peters and Dryden (2011) also noted in this study that graduate students are often responsible for backup of data and often leave the school with the data they collected and no one involved in the study knows where it is located. The students reacted positively when they were informed that the library was interested in providing data management services.

Each student enthusiastically supported the preservation and sharing of their data. It became clear from the interview responses that the workshop should stress the importance and the advantages of data preservation and sharing in the interest of knowledge advancement.

In 2011, Fear conducted research into how researchers manage their data at the University of Michigan. She discovered that faculty are not aware of what data management is and, in particular, of how or why they should share their data. The findings in the literature and the findings at OSULP are similar: data is varied, it is not being well-managed in some instances, and faculty and students are not aware that libraries can provide data management services. Yet librarians have many of the skills necessary to provide these services.

Similarly, after its purpose was described, the five students understood the importance of metadata for discovery and sharing. They also recognized the value of columnar, discipline-specific metadata and descriptions of equipment and methodologies that enable other researchers to understand and reuse data. The students especially understood the value of another scientist being able to reproduce their process. While it was assumed from the beginning of the project that metadata would be a large component of a data management workshop, it became apparent from discussions with students that an emphasis should be placed on teaching the role of metadata in making data reproducible and easily understood by scientists within their discipline and those in other disciplines.

Conclusions

As a result of the information gathered during the interviews and the literature reviewed, the team agreed to include the following topics in a new graduate student data management workshop offering: data storage and organization, metadata, data preservation, and data sharing. Existing curricula and data management workshops offered at other institutions were also consulted and further informed the content of the workshop. In order to garner enthusiasm from those required to attend the workshop, the benefits of data management are presented at the beginning of the workshop. Benefits include:

- Data management can increase the visibility and impact of the researcher and the research.
- It makes a researcher more efficient by avoiding digital archeology.
- It can further science cheaply and more efficiently.
- The integrity of research is better preserved if it can be replicated.
- And lastly, it may be a requirement of the funding agency.

During the interviews, students expressed a willingness to engage in good data management practices but generally lacked any understanding of how to go about it. The instructors therefore encourage class discussion about habits that can be developed and implemented while creating data and while managing and using data throughout the entire data lifecycle. For example, the value of applying conventional naming schemes is discussed and students are encouraged to use them from the beginning of the data-gathering process and consistently thereafter. Students are also encouraged to begin creating descriptive metadata while the data is being collected rather than after. This enables researchers to avoid the practice known as digital archeology. If the researcher describes data as it is being collected, there will be no need to find old notes or remember what a file name means. It will all be recorded in a logical and easy to access manner. Storing multiple copies of data in multiple locations throughout all stages of the data lifecycle is emphasized in the workshop. Students are advised never to alter raw data but instead to create a copy and alter that, leaving the raw data untouched.

The beginning of the workshop includes brief introductions during which each student introduces himself or herself, states the department in which they work, and describes the data with which they are working. This provides the facilitators with an opportunity to add content to the workshop on the fly that is specific to the attendees. The data management workshop has been conducted twice each quarter and once during each break. It generally has excellent attendance with 177 participants attending 12 workshops, averaging 15 per workshop. There has been participation from many colleges including Engineering; Agriculture; Earth, Ocean and Atmospheric Sciences; Business; and Forestry. There have also been participants from Student Health Services, Student Life, and other administrative departments. Participants include 14 faculty members, 155 graduate students, and two undergraduate students. Lab managers and administrative personnel have also attended in small numbers. Nearly every department on campus has been represented at these workshops. The participants have generally been willing to share how they work with data and participate fully in the workshop.

There has also been a positive participant response. One student was even overheard saying that she thought the workshop should be required for graduate students. The students also show their enthusiasm by asking questions during and after class regarding their specific data. Individual data consultations have resulted. Over time the workshop has been further developed to include an additional 40 minutes of workshop time (expanding the entire workshop to 90 minutes) and to include hands-on activities using metadata creation software.

Next Steps and Future Activities

One of the longer term goals for graduate student data management services at OSULP is to capture relevant graduate student research relating to theses and dissertations, data that is currently lost to the university and the world as soon as a student graduates. As a way to build on the information we learned in the student interviews and to better prepare for building a workflow and services for the capture of this data in repositories, OSULP is developing a graduate student survey. The survey is intended to help us better understand how many data sets are being created by students, in what types and formats, and how students are managing them. The survey will indicate the college of the graduate students so we may alter services by academic unit if that seems appropriate.

Since the initiation of this research, OSULP hired a Data Management Specialist to oversee the investigation and growth of data management activities. Data management workshops have been tailored for and presented to specific departments. The Data Management Specialist is teaching a graduate-level two-credit course on data management. Like the workshop, the course is cross-disciplinary because it was thought there would not be enough enrollment to have credit courses for specific disciplines. The course provides an opportunity to expand on topics that can be discussed only briefly in the workshop. For example, managing versions of files is discussed in the workshops, and several software packages are suggested as possible solutions. In the course, an entire session consists of teaching students to use and customize a particular open source solution. Another example is the use of metadata creation tools. In the workshop, several metadata creation tools are discussed and demonstrated. In the course, a two-hour session is devoted to the use of these tools, and each student uses one of these tools in describing their own data.

Assessment activities are being researched and discussed. While this research began by working with a small group of graduate students, the new services that have arisen from it are extensive and are gaining momentum. These efforts have led to many successful workshops, consultations, library publications, websites, outreach activities, and research ideas. More granting agencies are requiring data management plans and data is only getting bigger and more unwieldy. It is expected that data management services will continue to expand and develop at OSULP and other research institutions as librarians strive to meet the varied information needs of university faculty and students by adding value to their research in new and innovative ways.

References

Bicarregui, J., Gray, N., Henderson, R., Jones, R., Lambert, S., & Matthews, B. (2013). Data management and preservation planning for Big Science. The International Journal of Digital Curation, 8(1), 29-41. doi:10.2218/ijdc.v8i1.247

Boock, M., & Chadwell, F.A. (2010). Steps toward implementation of data curation services at Oregon State University Libraries. Retrieved from http://hdl.handle.net/1957/20593

Carlson, J., Johnston, L., Westra, B., & Nichols, M. (2013). Developing an approach for data management education: A report from the Data Information Literacy Project. The International Journal of Digital Curation, 8(1), 204-217. doi:10.2218/ijdc.v8i1.254

Carlson, J., & Stowell-Bracke, M. (2013). Data management and sharing from the perspective of graduate students: An examination of the culture and practice at the Water Quality Field Station. portal: Libraries and the Academy, 13(4), 343-361. doi:10.1353/pla.2013.0034

Collie, W. A., & Witt, M. (2011). A practice and value proposal for doctoral dissertation data curation. The International Journal of Digital Curation, 6(2), 165-175. doi:10.2218/ijdc.v6i2.194

Delserone, L. M. (2008). At the watershed: Preparing for research data management and stewardship at the University of Minnesota Libraries. Library Trends, 57(2), 202-210. Retrieved from http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1256&context=libraryscience

Duke University Office of Research Support. (2007). Research records: Sharing, retention, and ownership. Retrieved from https://ors.duke.edu/orsmanual/research-records-sharing-retention-and-ownership

Erway, R. (2013, December 60). Starting the conversation: University-wide research data management policy. EDUCAUSE Review. Retrieved from http://www.educause.edu/ero/article/starting-conversation-university-wide-research-data-management-policy

Fear, K. (2011). "You made it, you take care of it": Data management as personal information management. The International Journal of Digital Curation, 6(2), 53-77. doi:10.2218/ijdc.v6i2.190

Krishnankutty, B., Bellary, B. S., Kumar, N. B. R., & Moodahadu, L. S. (2012). Data management in clinical research: An overview. Indian Journal of Pharmacology, 44(2),

168-172. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3326906/

National Institutes of Health Office of Extramural Research. (2003). NIH data sharing policy and implementation guidance. Retrieved from http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

National Science Foundation. (2011). NSF data management plan requirements. Retrieved from http://www.nsf.gov/eng/general/dmp.jsp

Peters, C., & Dryden, A. (2011). Assessing the academic library's role in campus-wide research data management: A first step at the University of Houston. Science &Technology Libraries, 30(4), 387-403. Retrieved from http://dx.doi.org/10.1080/019426 2X.2011.626340

Steinhart, G., Saylor, J., Albert, P., Alpi, K., Baxter, P., Brown, E., Eli, Chiang, Kathy, Corson-Rikert, Jon, Hirtle, Peter, Jenkins, Keith, Lowe, Brian, McCue, Janet, Ruddy, David, Silterra, Rick, Solla, Leah, Stewart-Marshall, Zoe, Westbrooks, E. (2008). Digital research data curation: Overview of issues, current activities, and opportunities for the Cornell University Library. Retrieved from http://ecommons.cornell.edu/handle/1813/10903

Texas A&M University Libraries. (2013). Research data management. Retrieved from http://guides.library.tamu.edu/DataManagement

The University of Edinburgh Information Services. (2014). Research data management policy. Retrieved from http://www.ed.ac.uk/schools-departments/information-services/about/policies-and-regulations/research-data-policy

University of Minnesota Libraries. (2007). Understanding research behaviors, information resources, and service needs of scientists and graduate students: A study by the University of Minnesota Libraries. Retrieved from http://purl.umn.edu/5546

Witt, M., & Carlson, J. (2007). Conducting a data interview. Retrieved from http://docs.lib.purdue.edu/lib_research/81/

Appendix A—Interview Protocol

Questions for Students

Please note: This information will prove useful to provide a data management plan, but it is not expected that you will have answers to all questions. Any information you can provide is helpful.

General:

- Which degree program are you in?
- What is your research about?
- Are you collecting data? (or using a dataset(s) which is already available?)

Data description: General

- Describe the data you are collecting.
- How would you describe its relationship(s) to existing data?
- How much data are you collecting? (Units?)
- Can you describe the steps you anticipate from raw data collected to final analysis for publication?

Data description: Collection & Analysis:

- How are you capturing the data? What software are you using to capture, manipulate, and store the data?
- How are you structuring or organizing the data?
- In what format is the data? (e.g., Tabular? Spatial?)
- How/when will the data be processed and interpreted? What software or hardware are you using for data processing?
- How/where are you storing and backing up data files as you work with it? Will you keep different versions of the data?

Data description/Metadata:

- Is any metadata being created automatically as the data is captured? For example, cameras often capture the time and date along with a digital image.

- Are you aware of metadata standards in place for describing data relating to your topic?
- What other metadata do you think might be needed to describe this data so that it is understandable to others?
- How is the metadata being created?

Data dissemination/ownership

- Who/what has responsibility for safekeeping of this data?
- Who owns this data?
- Are there requirements to share this data?
- Are there privacy issues with this data?
- Are there copyright issues with this data?

Data preservation and Accessibility

- Who else might be interested in this data over time? (You, your research group, funder, other researchers, the public, policy makers?)
- For what length of time will this data have value to other researchers?
- Are you aware of any discipline-specific data repositories that might be a logical place to archive your data and make it available?
- Would you be willing to put your data in ScholarsArchive@OSU alongside your completed thesis/dissertation?
- Would you be willing to put a link to your data from your thesis/dissertation in ScholarsArchive@OSU or the catalog record for the item in ScholarsArchive@OSU?
- For how long should the data be stored?

Attribution:

- How would you like your data to be cited?