2013

Data Management and Sharing from the Perspective of Graduate Students: An Examination of Culture and Practice at the Water Quality Field Station

Jake R. Carlson
Purdue University, jakecar@umich.edu

Marianne S. Bracke
Purdue University, mbracke@purdue.edu

Data Management and Sharing from the Perspective of Graduate Students: An Examination of Culture and Practice at the Water Quality Field Station

Jake Carlson and Marianne Stowell Bracke
Purdue University

Abstract: Libraries are actively seeking to identify and respond to the data management and curation needs of researchers. One important perspective in this area that is often overlooked is that of graduate students. This case study uses the Data Curation Profile Toolkit to interview six graduate students working for Agronomy researchers at the Water Quality Field Station (WQFS) research lab to understand the students' practices with data, the challenges they face, and their attitudes towards managing and sharing data. Though a small study, this research could provide new insights for libraries creating data services, particularly in regards to graduate students.

Introduction

There has been a great deal of interest by libraries in developing services to support data management needs and make research data more accessible as a normative part of scholarship.[1] Constructing effective services requires a thorough understanding of current practice with regards to the handling, administering and the application of research data in the research process, as well as the motivations, resources and needs of the researcher.[2] Without such an understanding researchers are unlikely to make use of the services provided.

This paper depicts a project conducted by the Purdue University Libraries to identify barriers in managing and sharing data at the Water Quality Field Station (WQFS) at Purdue University through an examination of the perspective and practices of the graduate students who work there. As is the case in many science laboratories, graduate students at the WQFS have a significant role in collecting, processing and analyzing research data. The decisions they make and the actions they take in administering this data as it proceeds through its lifecycle are likely to have a significant effect on the quality and usability of the data for the lab. Moreover, the ability to share data with affiliates and others in ways that they can understand and use rests upon how well the data are documented, described, and

organized, tasks that are typically done by graduate students over the course of their work. Understanding the perspective and practices of graduate students is an important precursor towards facilitating effective data sharing.

Background

The Water Quality Field Station comprises 991 acres of land in Northwestern Indiana and has been in operation since 1992. The research conducted at the station focuses on developing practices to minimize the amount of chemicals from fertilizer or other sources from entering water supplies. These practices are evaluated from environmental, agronomic and economic perspectives with the intent of developing ecologically-balanced technologies for crop production. The WQFS generates multiple data streams such as water quality, water flow through tile drains, soil quality, measures of plant yields and genetic composition, and other factors relevant to the environmental impacts of agricultural practice and land use.

As is the case with many small labs, the WQFS lacks a robust infrastructure for managing their data effectively. Compounding the problem, the number of data streams and the amount of data collected and managed by the WQFS has grown dramatically over the years. Numerous scientists, staff, and students have participated in the collection, handling, management and storage of these data sets. The scientists and students each have had their specific interests in working with the data and have handled the data accordingly. The data generated at the WQFS fall into what Bryan Heidorn described as "dark data".[3] Dark data are data sets that are "not carefully indexed and stored so they become nearly invisible to scientists and other potential users and therefore are more likely to remain underutilized and eventually lost".[4]

The principal investigators (PIs) overseeing the WQFS have expressed a desire to do more with their data through making it more accessible, but are not sure how to go about doing so. Their situation is

similar to other "small science" labs identified in the literature, including: experiencing confusion around how and when to share their data, recognizing the need for standardized terminology and practices with their data, and an interest in having help managing, describing, and archiving data.[5,6] The WQFS has received numerous requests for access to the data over the years. These requests often come from researchers who use or design models to simulate the effects of management decisions or environmental change on water, soil, crops, etc. Making research data available outside of the lab in which it was generated is not yet a common practice in agronomic fields, and so there are few known frameworks for the WQFS to follow. The Purdue University Libraries has worked with the PIs on previous occasions to help address some of the issues in managing their data in ways that enable its dissemination outside of the WQFS.[7]

As a part of this effort, the authors attended a meeting of graduate students working with data generated in the WQFS labs. From our previous interactions with the PIs, we had an understanding of the issues in sharing this data externally, and we believed that talking with the graduate students would extend our understanding of their situation. At this meeting, students articulated some of the issues, questions and frustrations they had experienced in working with and sharing their data. These issues extended to sharing data with locally situated personnel who were affiliated with the WQFS and even within the WQFS group itself. The discussion at this meeting indicated that a more thorough examination of data management and local data sharing practices amongst graduate students could be beneficial to address the larger data management and sharing needs of the WQFS labs as a whole.

Literature Review

Librarians and others have conducted multiple explorations into researcher needs and behaviors for data to develop the foundational underpinnings for providing data services. These explorations have

been conducted at a variety of scales, and frequently center on identifying significant issues or commonalities across the researchers being studied.

Institutional level investigations for example typically seek to identify areas of need that are common to researchers situated at a particular location regardless of their discipline. The University of Minnesota conducted a wide ranging investigation into the information practices, behaviors and needs of researchers and graduate students which included discussions on data practices. The lack of guidelines and policies surrounding the organization, description and retention of data was found to be a significant issue, as was data storage and security.[8] Brian Westra at the University of Oregon interviewed 25 scientists and found similar issues with data storage and a lack of formal policies for administering data. Westra also noted an expressed desire by researchers for data management tools and resources.[9] More recently, a study done by the University of Houston Libraries found that researchers needed more assistance with meeting the new data management requirements of funding agencies from librarians than with data storage.[10]

Other studies have taken the opposite approach, examining practices and needs of researchers from a particular discipline across multiple institutions. In 2009, the Research Information Network (RIN) released the results of a study that sought to chart the information practices and exchanges of life scientists, including research data. They identified significant differences between the information exchange practices within the fields of study of the life sciences. RIN also found a disconnection between researcher behavior and the policies and strategies of support agencies designed to assist researchers in using and exchanging information.[11] The Digital Curation Center (DCC) conducted a series of case studies to investigate the attitudes and approaches of researchers across multiple disciplines towards data deposit, sharing, reuse, curation and preservation with an objective of identifying good practice in these disciplines. A major finding from this study was that the diversity of data types and

tools, as well as the skills, methods and needs of the researchers in working with their data meant that disciplinary level examinations, even within very specific fields of study, are not an effective approach to identify practices and requirements of researchers. Instead, this type of information should be sought at finer levels of examination such as the research groups themselves.[12]

Investigations of researcher needs have also been conducted on larger scales. Two examples are the initiatives launched by SURF and DataONE. SURF, a consortium of Dutch universities, conducted a review of the literature on research data storage and access to provide a collective summary of findings that could be used as a foundation for launching data initiatives. Amongst their findings was the need to support the day to day work of researchers, not just curation activities, and for researchers to retain control over what happens to their data.[13] DataONE's research team conducted a wide ranging survey spanning institutions and disciplines on researcher's perceptions and practices in data sharing. Amongst the many reported findings is that a lack of time and insufficient funding are significant barriers to researchers sharing their data. The survey also recorded differences in practice and attitudes by age of the researcher, geographic location (continent) and discipline.[14]

A significant gap in efforts to understand the practices of researchers through case studies, surveys or other means of investigation is the overall lack of attention given to the role of graduate students and their work in generating, processing, analyzing and managing data. Although the faculty researcher is the driving force and the intellectual authority behind the research, many faculty and labs rely on graduate students to perform day to day tasks that are needed to conduct the research and generate results. Given their intimacy to the data, the perception and attitudes of graduate students towards data management issues and the actions they take (or do not take) through the data lifecycle are likely to have a sizable impact on the later deposit, sharing and curation of a data set. If libraries and other support organizations are to develop effective services that address the real world needs of research

communities and labs, then developing an understanding of the approaches and practices of graduate students towards the data they work with is essential.

In addition, as the future researchers in their discipline, acculturating graduate students with sound data management practices at this stage is an important consideration. Research conducted at Purdue indicates that faculty see the knowledge and abilities of their graduate students with data as lacking and that there is a need for educational programing to teach these skills.[15] To meet this need several initiatives to educate graduate students in this area have been launched[16,17,18] and many more are likely to follow. Successful educational programs will be based on not only an articulation of what students need to know, but an understanding of current practice and the environments and cultures in which they work. Graduate students are not only learning the subject knowledge needed to be successful, but they are learning to become part of the community of scientific researchers. The lab setting provides a mix of peers and faculty mentors in a less formal setting than a classroom. This allows for the transfer of disciplinary norms and practice (in this case new norms and practices) both laterally and from experts to novices.[19] Furthermore, studies show that the social interaction and real-life application of the lab setting can be me even more important in the educational process than classroom work.[20] Therefore knowledge of the lab environment from the graduate student's perspective must be a consideration for planning educational programs and other data services.

Methodology

This project was carried out in three stages. The first stage centered on collecting information from graduate students about their data management and sharing activities and their opinions on these activities through in-depth interviews. In the second stage, the interview data were reviewed to determine the practices of individual students that were significant in managing or sharing their data. A second review identified areas of need or concern expressed by multiple graduate students in the

interviews. These practices and needs were then analyzed to identify any connections or relationships between them. The identified connections were categorized as a means of highlighting the high-level needs expressed by the graduate students. Finally, we responded to the categories of needs by generating recommendations for taking action to address them. Our results and recommendations were written up as a report on delivered to the directors of the Water Quality Field Station. Subsequently, we have met with the directors of WQFS on several occasions to discuss the contents of the report and to initiate action on some of the recommendations.

This project began in the summer of 2011. In the first stage of the project, the authors interviewed these six graduate students using a modified version of the Data Curation Profile (DCP) Toolkit developed by the Purdue University Libraries, based on research done at Purdue and the University of Illinois at Urbana-Champaign.[21] The DCP Toolkit is comprised of a semi-structured interview protocol in which researchers are asked questions regarding a particular data set that they are developing. The Toolkit also provides a structured framework for summarizing the researcher's responses and representing them in a way that allows comparisons to be made across multiple DCPs. A completed DCP will contain three elements. First, it will contain detailed information about the particular data set discussed in the interview. This includes information about the lifecycle stages of the data set and the characteristics of the data at each stage. Second, a profile will contain information about the current practices of the researcher in administering and managing the data set. Third, the profile will include information about what a research sees as problem areas or as unmet needs for managing or curating his or her data set.[22] This project was reviewed and approved by Purdue University's Institutional Review Board. All participating students gave their consent to be interviewed, had the opportunity to review the Data Curation Profile generated from their interview, and gave their consent for their DCP to be published.

The six graduate students interviewed represented a cross section of the research being conducted at the WQFS. Student research includes how a particular species of switch grass (as a biofuel) processes nutrients, the effect growing bioenergy crops has on soil structure and quality, and how changes in the types of crops grown on a particular plot of land impact the amount and nutrient quality of the water in the sub-surface drainage. Most of the students are formally enrolled in the Agronomy department, but one is studying Agricultural and Biological Engineering. With the exception of one student working towards his Master's Degree, the students were all pursuing PhDs.

The standard data interview for the DCP consists of 13 modules and is meant to be conducted over two sessions. Due to the schedule of the graduate students we decided to reduce the number of modules and conduct the interview over one session instead of two. In addition to the four required modules of the DCP: "data set", "data lifecycle", "data sharing", and "organization and description", we also included two of the optional modules: "tools" and "data management". As the interviews were semi-structured in nature, the students would sometimes bring up issues outside of those addressed by the interview questions; most notably intellectual property issues.

Once the interviews had been transcribed, the authors generated a DCP from each of the interviews. The six DCP were reviewed for statements in which the graduate student expressed a need or indicated an issue or concern in working with their data. These statements were extracted and compared across the participating graduate students. Statements that expressed a similar issue or need were grouped together for further analysis. In addition, issues that were deemed to be of particular significance to one of the graduate students were also included in the analysis, even if the issue was not discussed by the other graduate students. These statements were then examined as a whole and grouped into fifteen areas of concern. The fifteen areas were then grouped once more into four over-arching

categories: “Data Documentation and Organization”, “Data Sharing”, “Long-Term Data Management” and “Ownership and Authority over the Data”.

Results

These four categories, while overlapping, highlight the distinct areas of concern for the Water Quality Field Station as it works towards developing better local data management and sharing practices.

1. **Data Documentation and Organization**

Data documentation and organization was a significant topic of discussion in the interview and a source of concern for many of the graduate students. Overall, students reported a lack of clear and shared expectations as to how data should be documented and organized in the WQFS. Although there is interest in sharing data from the WQFS with others outside of the research center, there has not yet been action taken to articulate what it would take to enable their data to be usable by others, or steps taken to create guidelines or requirements in working with data. This is not to say that the students did not receive any instruction or guidance. Their advisors provided direction to students during their frequent interactions. However, these interactions tended to be limited in their scope, focusing on the more immediate issues at hand rather than a longer-term, bigger-picture view of effective data management and sharing. In the absence of formal data management plan to follow, the graduate students developed methods of documenting and organizing their data based on discussions with their advisor and a sense of what is needed for their specific purposes. As one of the interviewees stated:

> “[G]raduate students really don’t have the means or the, in terms of time or length, perspective that faculty do. So they’re going to implement what they’re told to implement, functionally. I mean we’re very creative, we’re happy to create and innovate and make new things, right? But we’re doing that in our research, so when it comes to maintenance or like housekeeping of the

> research really everybody's work is such a small sliver that they don't have that bird's eye view… They'll come up with [documentation] for themselves, which means it's not going to make any sense to the next [student]… that's the problem."

Students also reported that their approaches were largely informed by their previous experiences or training in working with data, to the extent that they had any. As would be expected in this situation, the degree and the depth to which students documented and organized their data varied widely from student to student. Some students had developed their own standard operating procedures for their data, although they confessed that they did not feel completely successful in this endeavor. Others created their own reference materials, such as a data dictionary or a master spreadsheet, to assist themselves and their advisor in working with the data.

A significant obstacle cited by students is the lack of known and agreed upon standards for managing, describing, organizing and sharing data in Agronomy and related fields. The absence of such standards makes it difficult for faculty and for students to know where to begin in crafting their own standardized approaches at a local level in the lab or for themselves individually. One student did report a growing recognition that such standards are needed and that groups within her field are forming to work on this issue.

1.1 – Handling and Use of Lab Notebooks

One common element across the students was their use of a lab notebook to document and organize their work. The information entered into their lab notebook served several purposes. One student referred to her lab notebook as her "work diary", as it contained a daily report of her activities on any given day, even if that activity was as mundane as cleaning the lab equipment. Like many students, she uses her lab notebook to back track and check her work if something unexpected appears in her results. Her lab notebook also contained information that supported her work routines, including her "recipes"

for processing her data, references to the protocols that she used, and any other methodological procedures that she follows. Finally, she prints out copies of her data (spreadsheets and images primarily) and tapes them into her lab notebook. She will then label and annotate her data in the notebook as the primary means of documenting her work. The other graduate students followed a similar protocol though the extent of the information entered into the lab notebook varied. For example, not every student included information about the source of the data print outs that were added into their notebooks. Instead, they rely on the units of measurement and other distinguishing characteristics of the data print out to identify the equipment used to generate the data and/or the source file containing the data described in their lab notebook.

All of the students cited their lab notebooks as the definitive source of information about their data, and several students stated that the information contained in their lab notebook would be needed if their data were to be shared with other students. However, as several students noted, their lab notebooks in their current form are of limited use in enabling data to be understood by and shared with others. Some of the students only document a part of their data lifecycle in their lab notebooks. Once they progress into data analysis, or even data processing, documentation is done through other means, mostly electronically. A common method of description stated by the students is to make annotations in their spreadsheets alongside the data itself to explain a particular variable, define the process used to generate the variables, or indicate any deviations from expected results. Furthermore, as one student observed, lab notebooks do not necessarily provide a complete picture of their work. As an Agronomist, this student is employing the methodologies common to Agronomy in developing his data set and is making certain assumptions in documenting (or not documenting) portions of his work as a result. These assumptions are generally understood by his advisor and likely to be understood by those doing similar types of research, but if his data were to be shared outside of his particular research focus, he would need to provide a more detailed explanation of his work.

Even in cases where a sufficient amount of information about the data for others to understand and make use of the data, the structure of the lab notebooks hinders the utility of the documentation it contains. The physical nature of the lab notebook makes it difficult to share its contents with others, even those who are located in close proximity to each other. Lab notebooks are not easily copied and lending them out is impractical. Entries in lab notebooks are typically chronologically based which can make locating specific pieces of information a challenge.

1.2 – Handling and Use of Electronic Data Files

The students interviewed generated a great deal of data over the course of their research. The lifecycles of their data varied from project to project but generally followed a common pattern of collecting data from the field, processing the data to generate usable variables, analyzing the data and then publishing elements of the data through generating tables and charts for presentations or to add to papers. Although data may be captured and analyzed in different formats by students over the course of the data lifecycle, Excel spreadsheets are their format of choice in managing the data. Students varied in their approach to organizing their data within Excel. Several students placed their processed data in the same spreadsheet as their raw data, making distinctions between raw and processed variables through column headings, annotations or using separate tabs. Annotations were also frequently used to note anomalies in the data or in explaining places where the practices used to gather or process the data deviated from the norm in some fashion. In discussions with the Principal Investigators of the WQFS about the findings of this study, they reported that the information entered into these spreadsheets as annotations or comments is usually insufficient for them to understand the students' data fully. The low learning curve and basic functionality in handling variables make Excel an appealing tool for housing scientific research data, however its lack of descriptive and organization capabilities present challenges for those seeking to understand and make use of such data.

Over the course of the data lifecycle, students generated multiple files for the purpose of testing and analyzing the data. Students did not follow any one set practice in managing the files they generated; instead organizing the data within these files according to what would be useful for their particular purposes and needs. A notable exception is the student who reported that his faculty advisor requires him to keep a Master Spreadsheet that serves as an official record of the data once it has been tested by the student and accepted by his faculty advisor. In sum, the description and organizational frameworks developed by students' were generally geared to meet more immediate needs rather than to support mid to long-term usage of the data.

2. **Data Sharing**

Making data available for others to view or use has received considerable attention from funding agencies, scholarly societies, information scientists and librarians, and open access advocates. However in practice, the act of sharing one's data with another is not well recognized or supported in Agronomy as a part of normative practice.

2.1 – Attitudes toward Sharing Data

In speaking with the graduate students at the Water Quality Field Station, we found that students were generally open to sharing their data with others, under certain conditions. Students spoke highly of sharing data in abstract terms, stating that it is a part of good scientific practice, data gathered using public funds should be made available to the public, and data sharing would further the aims of their research communities. Speaking at a more personal level, one student stated that having her data associated with her publications would improve people's ability to understand her work.

However, students did express some concerns. The timing of the release of their data was an issue for most of the students we interviewed; they wanted the opportunity to publish the results of their work

prior to making the data available. One student highlighted her concern through stating that releasing her data before completing her publications would lead to her feeling pressured to publish before she would be ready to do so. Another student stated that she would like to be contacted first, so that she could understand the intentions of the individual requesting the data.

Several graduate students expressed some trepidation over the potential for others to misunderstand or misuse their data. One student spoke very highly of her fellow students and indicated that she would be willing to share her data with them as she trusted that they would not misuse her data. However, the trust she expressed in her colleagues did not extend beyond the WQFS, and so she was not altogether comfortable in making her data more widely available. A particular issue raised by the students in consideration of sharing their data was uncertainty over who would be using their data and for what purpose. In contrast to their general support of sharing research data, it was difficult initially for many of the students to conceptualize what value their data might have for others; though after further consideration, some of the students were able to articulate specific applications in research or in the field.

2.2 – Data Sharing at the Disciplinary Level

One possible real-world situation for data sharing that was discussed during the interviews was the application of the data generated at the WQFS to the computer models developed or refined by researchers in Agricultural and Biological Engineering (ABE). These computer models employ algorithms to make predictions about the effect of land management decisions on elements such as water quality, sediment, soil composition and crop yields through the application of a variety of input variables pertaining to land use and agricultural practices. Researchers who engage in modeling require data to demonstrate the validity of their model. The Principal Investigators have been generating data sets over many years at the WQFS that would be relevant for this purpose, and they are often contacted by

modelers asking for the data. However, ABE modelers employ a different set of assumptions and approaches in their work than agronomists do. This disconnection between different fields of study and their intended uses for the data are a serious impediment to sharing data effectively.

One of the interviewees was a graduate student in ABE who employs models as a part of her work. She described herself as straddling two worlds as a producer of agronomy data and as a data user with her work in computer modeling. The challenge as she sees it is for the data producers is to not only include information about the context of their data into the documentation, but to convey this contextual information in ways that could be understood by the likely consumers of the data, modelers in this case. Modeling work requires that uncertainties in the data have been identified and addressed to enable the modeler to present her work with a high degree of confidence. Deviations in the data or variations in the data over time introduce uncertainty into the work of the modeler and so documentation is needed to explain these differences in the data from what was expected. However, without the associated documentation and other descriptive information providing this contextual information it is difficult if not impossible to represent the data faithfully in the model.

2.3 – Data Sharing at the Local Level

One of the Agronomy students in his interview recounted an experience in sharing his data with a student in Agriculture and Biological Engineering (ABE) at Purdue (who was not interviewed for this project). The student in ABE asked for the Agronomy student's data on plant lignin to test a model he was working on. The Agronomy student sent him several Excel spreadsheets that contained his data; however once the ABE student began to examine the data several questions arose. The ABE student did not understand some of the data points in the spreadsheet, which led to several email exchanges between the students to try and understand the discrepancies between what the ABE student expected and what the Agronomy student had delivered. Eventually the ABE student consulted with his professor

and together they realized that the assumptions used in generating the data in an Agronomy lab were different from what would be used by ABE researchers in testing their models. This interchange between students demonstrates the challenge of sharing data that was generated for a particular purpose with another researcher who wishes to reapply the data for a different purpose. Although local proximity and access to the data producer facilitates the act of sharing data itself, it does not negate the need for sufficient documentation and description to ensure that the data consumer can understand and make use of the data.

Other obstacles described in the interviews were the uneven levels of awareness about what data sets were being generated or managed by others at the WQFS and uncertainty over the protocol in requesting access to these data sets. One student mentioned that she is potentially interested in obtaining data from some of the other students in her lab as their work may augment the research she is engaged in is taking place in the lab. However, she hesitates in pursuing her interest in this data because she is uncertain who to ask about it, she is unaware of how far along they are in developing the data set, and she does not know how much time and effort it would take to locate the specific portions of the data that she would be interested in and prepare it for her purposes. This student did mention that graduate students affiliated with the WQFS had recently started holding meetings to share information about their research and that these meetings were quite helpful in giving her a window into the data being generated. Given these statements, it appears that the levels of awareness of the data being generated locally at the WFQS varies from student to student and depend mostly upon the social networks of the student. The WQFS station is comprised of multiple teams, each with their own research agendas. In this environment, it is not surprising that graduate students may not possess more than surface-level knowledge of the data being generated by other teams.

2.4 Lack of Models and Structures for Sharing Data

The students interviewed reported that sharing data publicly is a rarity in Agronomy and related fields. There are no large data repositories that serve as community resources. Furthermore, most of the students interviewed reported that the journals they expect to publish in do not accept data as a supplementary file for publication. Even the one student who did state that some of the journals she would consider submitting her work to would accept supplementary data files stated that it was not a common practice. Despite this situation several students expressed an interest in associating their data with their eventual publications. One student expressed disappointment that publishers did not accept data files, stating that their inclusion would give her more confidence in the research being presented:

> "I think it's really important and helpful because if I had other people's data I could directly compare instead of saying, well, they assumed and I kind of think that's what they meant, but I don't really know."

Most of the students expressed a belief that the information presented in the journal article would generally be enough for someone else in their field to understand their data.

> "… the publication is the concise representation of your research and in identifying the trends or things that are of significance and placing that within the context of the greater body of knowledge. So with the publication [it] would have everything explicitly outlined, how each experiment was done or at least link to an explicit explanation."

Several of the students felt quite strongly that if their data were to be made available to others, that the data should be linked to the publication in some manner so that the consumer could access the data from the publication and vice-versa.

The lack of resources or structures in the lab and in the discipline to support sharing data makes it difficult to overcome the current status quo. The scarcity of models or best practices that students

could draw from likely limits their recognition of the potential value of their own data as an information resource and their understanding of how to construct and document their data in ways that would aid sharing it with others.

3. Long-Term Data Management

Several issues pertaining to the long term care and management of the data came to light during the interviews. These issues include the practice of inheriting data sets, the lack of an infrastructure to maintain these data sets, and some security concerns.

3.1 - Data Inheritance

A common situation faced by graduate students in the Water Quality Field Station is inheriting a data set that was crafted by a student who preceded them. The interviewed students reported varying degrees of success in working with data sets that they inherited. Generally, the data sets they inherited did not contain much in the way of description or documentation. Instead, their advisor, as the intermediary in the transfer of the data to the student, served as the primary, if not the only, source of information about the transferred data set. One student stated that his use of an inherited data set was facilitated by his advisor requirement that all students keep a master file of their data to serve as an official record of their work. Although information about the methodologies used to generate and process the data were not included with the data set, the student was able to make sense of the data due to the shared organizational structures employed by both students.

However, other students reported instances where their lack of familiarity with the data limited its utility or presented challenges for them. One of the students reported that she often seeks to make use of data points generated by others through integrating them into her own data sets; however she faces several challenges in doing so. First, there has been little uniformity in how data sets are organized and

documented as different students employ different approaches. Second, she is often unable to identify who in the lab was responsible for the generating or modifying the data points that she is interested in using. Another student discussed a situation in which she needed water flow and quality data that had been generated in the WQFS. She tracked down the data set, but the student who had inherited the data set stated that she did not have enough of an understanding of the data to be able to identify the precise aspects of the data that were requested. The student requesting the data eventually found what she was looking for in the data set; however it became apparent to her that some of the data had been manipulated and modified by others. No clear record was made of what modifications were made, when, or by whom, which made the data difficult for her to trust.

3.2 – Lack of an infrastructure to maintain digital data sets

The digital data generated by the WQFS has longitudinal value to the students and researchers associated with this facility. The majority of the students reported wanting to keep copies of their digital data sets, or at least the elements of their data that they considered to be important, as historical records of their work. However, none of the students had really given the long term maintenance of their digital data sets much thought or taken action to ensure long term access to their data. This is in contrast to the physical data samples that were collected. One student described the process that she has used to document and store the soil samples that she had collected at various points along her research. She has crafted spreadsheets that "profile" the samples, which includes the plots they came from, the depth at which the sample was collected, what was done to the sample in processing it, etc. When she graduates she intends to print out her spreadsheets and attach them to the samples themselves so that others in the lab can make informed decisions on how they might conduct further analysis on the sample, or determine if the sample is worth keeping at all.

3.3 - Data Tracking and Security

As the primary handlers of the data generated and analyzed at the WQFS, students are also serving as the defacto care takers of the data. The interviewed students were generally not proactive or very much engaged in taking action to ensure that good management or security practices were being followed. Graduate students generally assumed that the computing resources provided by their academic department were adequate and that the security of the data was being addressed by their IT unit. In contrast, one student recounted an incident when a faculty member she was working with made an accidental keystroke and deleted a file. When they contacted their IT unit to restore the file, they learned that the back-ups had not been taking place and so the data was lost. Students did report making back-ups of their data in several different fashions. Some students used their personal computers, others purchased external hard drives, and some chose to email their data files to themselves as their means of backing up their work. The frequency of their backups varied, but they were typically performed manually by the student when they believe that they have made significant progress.

The interviewed students stated that they kept earlier iterations of their data files to enable them to retrace their steps if they needed to do so. Typically, their system for distinguishing between iterations of their data was to indicate the version through its file name, often by including the date, or through placing the file in a particular folder. Students reported that their approaches tracking previous versions of their data were satisfactory for their purposes, although several students expressed some concern over their ability to maintain an overall accurate history of their work.

Over the course of their research, students reported that they needed access to their data at times when they were not in their office or could not reach their account on the secured campus network. In these situations, students will make working copies of their data for use on their personal computers. Students did state that they tried to be careful to reconcile the original data living on the university

network space and the modified files that were moved to their personal computers. However, they recognized that possessing multiple iterations of the data in different locations was not ideal and presented problems in keeping track of their work.

4. **Ownership and Authority over the Data**

Ownership over the data has not been formally articulated at the Water Quality Field Station. The interviewed students did report that they feel their data is not really "theirs" at all, but that it belongs to their advisor and the WQFS lab. Questions about ownership and authority over the data are complicated further by the frequent occurrence of students working on data sets that they did not originally generate themselves, but inherited from other students. Given this environment, it was not clear to the graduate students how much decision making authority they had, if any, over the data. The lack of clear statements over what students could or could not do with the data they were working with appear to act as somewhat of an inhibitor to students. As the students do not perceive themselves as having decision making power over the data it is not clear that they feel that they have much incentive to do more with the data other than to satisfy their immediate and individual research goals.

Discussion

The interviews with the six graduate students at the Water Quality Field Station revealed areas of concern in the documentation and organization of data, data sharing both within and outside of the WQFS lab, long term data management practices and ownership and authority over the data. Once the interviews were completed and reviewed, we discussed with the Principal Investigators of the Water Quality Field Station and generated a report that included a list of recommendations to address these areas of concern. Although our recommendations were necessarily targeted to the WQFS, they may well have applicability to other research labs in the Agronomy field or beyond.

1. High-Level Issues

At a high level, the current practices surrounding the data sets appear to be constrained by a couple of factors. First, although the PIs are very interested and active in seeking to do more with the data generated at the WQFS, a "lab culture" that would support their interests has not yet taken root. The interviews demonstrate that the handling and administration of data sets is primarily driven by the needs, perceptions, and skills of the individual who is in current possession of the data, the graduate student. The faculty advisor can, and often does, play an influential role in shaping the treatment and disposition of the data set, but here too the interactions between the student and their advisor generally take place in a localized context. Discussions between advisor and student on data management issues are generally held at the point of need, when the organization, documentation or other management issue is interfering with the ability of the student, advisor or others in the lab to make progress in their research. The forces driving the attention and actions taken by the advisor and the student with regards to the data are naturally centered on generating and extracting immediate or near-term value for the research being conducted. Considerations for later access or re-use by others in the WQFS lab are a secondary issue.

Second, echoing the statements made in the graduate student interviews, the PIs feel that their efforts to bring about a change in culture and practice at the WQFS is hampered by the absence of a larger disciplinary culture that supports the sharing and reuse of data. It is difficult to move in the direction of doing more with the research data being generated at the WQFS if the professional societies, journal publishers, and other researchers in the field do not offer support or incentives for doing so. In talking with the PIs, they see this lack of support primarily stemming from a lack of awareness and understanding in Agronomy and related fields of the potential benefits of making their research data more available. Agronomy is not a "big data" field and relies mostly on comparatively smaller and more

locally targeted funding. Without visible models or high profile examples it is difficult to know what actions to take or where to begin.

2. Overall Recommendation for the WQFS

In considering our recommendations, we recognized that any course of action that we proposed would need to be aligned with current practice and accepted norms in the WQFS to be successful. Although direct examples and models for managing and sharing data are scarce, researchers in the Agronomy field are used to following Standard Operating Procedures (SOPs) in their field work and research. Documenting and following SOPs are a vital part of research and practice in Agronomy as some of the equipment used can injury or even kill if it is used improperly. SOPs are also used to refer to the research methodology that is being employed. Thus, the list of recommendations that we crafted encourage developing SOPs for handling, documenting, sharing and managing data for the long term.

In analyzing the interviews of graduate students, we felt that the primary issue behind many of the concerns was the lack of defined and shared expectations for handling, documenting, sharing and preserving the data generated at the WQFS. Our overall recommendation for the WQFS is to take the information presented in our report and use it as a means to launch discussions with the intended outcome of determining the policies, practices and "lab culture" needed to administer their research data. We provided a possible approach for the WQFS to follow to craft a comprehensive solution to the data issues they face. We designed this approach to try and bolster awareness of the issues facing the WQFS with regards to their data through discussion at multiple levels in their organization and to encourage action. The basic structure of the proposed discussions is to have senior administrators begin to articulate what they want to be able to do with the data. From there, they would determine their expectations for handling, documenting, sharing and preserving their data collectively, and then to identify gaps between these expectations and current practice at the WQFS. Discussions would then be

expanded to include graduate students and other WQFS personnel. The outcomes of these discussions would serve as the basis for developing policies and SOPs, testing them out, implementing them and finally training WQFS personnel on how to follow them. The comprehensive approach we proposed is available as a separate document in Purdue's institutional repository[23].

3. Targeted Recommendations

We recognized that a comprehensive approach to data management may require more time and resources than the WQFS would be able to allocate at the moment, and so we included additional recommendations that are more narrowly targeted to address particular areas described by the graduate students in the interviews. Our targeted recommendations included the following:

- Identify the common elements that need to be included in documenting the data generated across the WQFS. This may include things such as the people involved with the data (creator, processor, etc.), important dates (harvest, processing, etc.), the conditions under which the data were collected, the methodology and/or equipment used to generate or process the data, etc. Consider these commonalities across the lifecycles of the data sets generated at the WQFS and identify milestones as a means to inform the development of these elements.

- Designate an appropriate person (or persons) at the WQFS to assume responsibilities for developing high-level policies and procedures on documentation and organization of data, either for the WQFS as a whole or for specific research projects. Identify this person as a resource to whom graduate students and others can go to with questions or for advice.

- Devote time to discussion about data issues with graduate students. These discussions could range from formal training sessions taught by faculty, to informal discussion time between graduate students at meetings or over coffee.

- Work with the Purdue University Libraries or other agencies to develop training programs or skill sessions for graduate students.

- Investigate ways and means to replace or transfer the documentation currently captured in physical lab notebooks with electronic replacements in full or in part. E-Lab notebook products are likely too expensive and impractical to introduce at this time, but there may be other tools that could be adapted and adopted, such as: wikis, Microsoft One Note, Google Docs / Spreadsheets, etc. Investigate the digitization of existing (or future) lab notebooks in ways that would facilitate direct association and connections to the data sets they document for the purposes of making the data easier to understand, use, manage and curate.

- Create a directory of data sets that are/have been generated at the WQFS and make this directory accessible to the lab.

- Encourage graduate students to submit their data to an appropriate repository (such as the Purdue University Research Repository) before they submit the article to a publisher. Cite the data set in their article using the assigned Digital Object Identifier (DOI). Once published, link the data set in the repository to the article through the DOI.

- Develop a policy and/or S.O.P.s on data management and security issues. The policy and S.O.P. may include statements on the following questions:
    - Back-ups: when should backups be performed and how often? What are acceptable backup devices to use?
    - Under what circumstances is it acceptable to make copies of the data? How should data be reconciled between data files?
    - Under what circumstances is it acceptable to take data out of the lab? How should data taken out of the lab be reintroduced into the lab?
    - What constitutes the official record of the data? What information should be included in or associated with the official record?

- Articulate the intellectual property rights (and responsibilities) for graduate students and others who are generating data at the WQFS labs. This may include:
    - Identify the owner(s) and stakeholders (those with a vested interest in the data) of data sets generated at the WQFS labs.
    - Statements on the decision making process about the data.
    - Clarifications on who is permitted to make decisions as to the handling and disposition of data, and under what circumstances.
    - Clarifications on what actions graduate students may take with regards to the data they work with. For example, are graduate students permitted to take a copy of the data with them when they graduate?

- Consider developing a data preservation plan for data sets of high value. Work with the Purdue University Libraries / Distributed Data Curation Center (D2C2) and others to explore possible approaches to preserve data sets of high value from the WQFS.

Conclusion

As depicted in this and other studies, managing, sharing and preserving data in "small science" settings in ways that add value to the researcher and to the larger research community is a complex task comprised of multiple challenges. Graduate students are part of these multiple challenges, as data collectors and generators at a basic level, and as the future researchers. Thus, understanding their roles and perspectives can bring fresh insights into services libraries can provide. Making these changes will require shifts at the disciplinary level as well as the practices and norms of individual research labs. Increasing pressure from external organizations, such as funding agencies and journal publishers, on researchers to make their data more accessible and sustainable for the long term is raising awareness of the need to take action. Libraries are stepping up to fill this need through developing data management and curation resources and services.

Libraries and other interested parties have been conducting needs assessments and other explorations to understand current perceptions and practices of researchers to ensure that the services and resources being developed align with norms and expectations. However, generating a complete understanding of norms and practices in the lab requires a wider perspective than that of the researcher alone. Graduate students are the ones on the front line of generating, processing, analyzing and managing data. Their attitudes and actions will affect the ability of researchers to fulfill obligations or take advantage of developing frameworks to recognize data set as important sources of scholarship in its own right. At a high level, the findings of our study align with other studies that have been done in this area. The WQFS would benefit by developing plans and policies for their data, practices in

organizing, storing and data security are a concern, and the PIs have expressed a need to preserve their legacy data in ways that maintain or add to their value. However, including graduate student in our needs assessment of the WQFS has provided us with not just a wider understanding of the issues confronting research labs as they seek to respond to new requirements and take advantage of new opportunities, but a greater depth of understanding as well. Graduate students are not a marginal component, but rather an integral piece of the data management and curation process. Their perspectives and needs should be considered by libraries in developing services.

Instituting changes in research practice can be difficult for established researchers; however, this can also be an ideal opportunity to reach out to graduate students as the future disciplinary researchers. Graduate students are at a stage where they are forming their professional identity through their training and education. They are open to forming their own research norms and practices. Connecting with graduate students in lab or field settings, where they are developing their skills as a practicing researcher may provide an even richer learning environment to encourage the development of good data management practices than the classroom. Here then are opportunities for libraries to work with faculty on creating educational or other programs to plant the seeds of change in the next generation of researchers. The Data Curation Profiles that were created from our interviews are available on the Data Curation Profiles Directory website.[24, 25, 26, 27, 28]

Our work with the WQFS continues in several ways. After our report was delivered, the WQFS lab group hired a retired agronomist to help develop systems for documenting and organizing data sets generated by the WQFS to ensure their continued usability and to prepare them for deposit into PURR, Purdue's data repository. Along with a WQFS graduate student, we are serving as consultants for this initiative by identifying existing standards that could be applied towards WQFS data and providing guidance on what would be needed for curation. The WQFS is also part of a larger Department of

Energy grant. We have begun to meet with these researchers and graduate students to explore applying our findings to the data generated by this larger group.

This case study suggests some further areas for research. For instance, developing a greater understanding of the social practices in the transfer of knowledge and procedures from faculty mentors to graduate students would be very useful in suggesting ways that new data curation practices could be taught and adopted. Additionally, the targeted recommendations would be a useful starting point for working with other labs to identify current practices and data management needs. We anticipate continuing to explore issues surrounding data management and curation as they relate to graduate students associated with the WQFS and beyond.

References

[1] Tyler O. Walters "Data Curation Program Development in U.S. Universities: The Georgia Institute of Technology Example" *International Journal of Digital Curation*, 3, no.4 (2009): 83-92.

[2] Kathryn Lage, Barbara Losoff, and Jack Maness "Receptivity to Library Involvement in Scientific Data Curation: A Case Study at the University of Colorado Boulder" *portal: Libraries and the Academy*, 11, no.4 (October 2011)

[3] P. Bryan Heidorn, "Shedding Light on the Dark Data in the Long Tail of Science," *Library Trends* 57, No. 2 (Fall 2008): 280-299, doi: 10.1353/lib.0.0036.

[4] Ibid., 280

[5] D. Scott Brandt, "Librarians as partners in e-research," *College & Research Libraries News*, 68, no. 6 (2007): 365-367, 396.

[6] P. Bryan Heidorn, "The Emerging Role of Libraries in Data Curation and E-science," *Journal of Library Administration* 51, no. 7-8 (2011): 662-672, doi:10.1080/01930826.2011.601269.

[7] Marianne Stowell Bracke, "Emerging Data Curation Roles for Librarians: A Case Study of Agricultural Data," *Journal of Agricultural and Food Information* 12, no. 1 (2011): 65-74, doi:10.1080/10496505.2011.539158.

[8] Cecily Marcus and others, *Understanding Research Behaviors, Information Resources, and Service Needs of Scientists and Graduate Students: A Study by the University of Minnesota Libraries*, University of Minnesota Libraries (2007), https://conservancy.umn.edu/bitstream/5546/1/Sciences_Assessment_Report_Final.pdf.

[9] Brian Westra, "Data Services for the Sciences: A Needs Assessment," *Ariadne*, 64. (2010), http://www.ariadne.uk/print/issue64/westra.

[10] Christine Peters and Anita Riley Dryden, "Assessing the Academic Library's Role in Campus-Wide Research Data Management: A First Step at the University of Houston" *Science & Technology Libraries*, 30 (2011): 387–403.

[11] Research Information Network and the British Library, *Patterns of Information Use and Exchange: Case Studies of Researchers in the Life Sciences*, (Research Information Network, 2009), http://www.publishingresearch.net/documents/RINPatterns_information_use-REPORT_Nov2009.pdf.

[12] Key Perspectives, *Data Dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long Term Viability* (Edinburgh: Digital Curation Centre, 2010), http://www.dcc.ac.uk/sites/default/files/documents/publications/SCARP-Synthesis.pdf.

[13] Martin Feijen, "What Researchers Want," *SURF*, February 2011, http://www.surf.nl/nl/publicaties/Documents/What_researchers_want.pdf.

[14] Carol Tenopir and others, "Data Sharing by Scientists: Practices and Perceptions" *PLoS ONE* 6, no. 6 (June 29, 2011): e21101, doi:10.1371/journal.pone.0021101.

[15] Jacob Carlson and others, "Determining Data Information Literacy Needs: A Study of Students and Research Faculty," *portal: Libraries and the Academy* 11, no. 2 (2011): 629-657.

[16] Jian Qin and John D'Ignazio, "Lessons Learned from a Two-year Experience in Science Data Literacy Education," in *Proceedings of the 31st Annual IATUL Conference* (West Lafayette, Indiana: IATUL, 2010), http://docs.lib.purdue.edu/iatul2010/conf/day2/5.

[17] Mary E. Piorun and others, "Teaching Research Data Management: An Undergraduate/Graduate Curriculum," *Journal of eScience Librarianship* 1: No. 1 (2012): 46-50, http://escholarship.umassmed.edu/jeslib/vol1/iss1/8.

[18] Jake Carlson and others, "Data Information Literacy" http://datainfolit.org.

[19] John C. Weidman, Darla J. Twale, and Elizabeth Leahy Stein, "Socialization of Graduate and Professional Students in Higher Education: A Perilous Passage?" *ASHE-ERIC Higher Education Report* 28, no. 3 (2001), http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED457710.

[20] Robert A. Campbell, "Preparing the Next Generation of Scientists: The Social Process of Managing Students," *Social Studies of Science* 33, no. 6 (2003): 897-927.

[21] Michael Witt and others, "Constructing Data Curation Profiles," *International Journal of Digital Curation* 4, no. 3 (2009): 93-103.

[22] Jake Carlson "Data Curation Profiles Toolkit," Purdue University Libraries http://datacurationprofiles.org.

[23] Jake Carlson and Marianne Stowell Bracke "A Possible Approach Towards Addressing the Data Management and Sharing Needs of the Water Quality Field Station" DOI: http://dx.doi.org/10.5703/1288284315041.

[24] Jake R. Carlson "Agronomy / Biofuels - Purdue University," Data Curation Profiles Directory: Vol. 3, Article 3. (2011) DOI: http://dx.doi.org/10.5703/1288284314991.

[25] Jake R. Carlson "Agricultural and Biological Engineering / Eco-Hydrology - Purdue University," Data Curation Profiles Directory: Vol. 3, Article 2. (2011) DOI: http://dx.doi.org/10.5703/1288284314990.

[26] Marianne S. Bracke "Agronomy / Grain Yield - Purdue University," Data Curation Profiles Directory: Vol. 3, Article 4. (2011) DOI: http://dx.doi.org/10.5703/1288284314992.

[27] Jake R. Carlson "Agronomy / Soil Microbiology - Purdue University," Data Curation Profiles Directory: Vol. 3, Article 5. (2011) DOI: http://dx.doi.org/10.5703/1288284314994.

[28] Jake R. Carlson "Agronomy / Land Use - Purdue University," Data Curation Profiles Directory: Vol. 3, Article 8. (2011) DOI: http://dx.doi.org/10.5703/1288284314993.