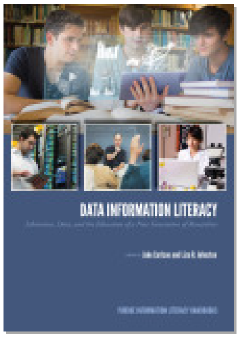
Data Information Literacy

Carlson, Jake , Johnston, Lisa

Published by Purdue University Press



Carlson, Jake and Lisa Johnston.
Data Information Literacy: Librarians, Data and the Education of a New Generation of Researchers .

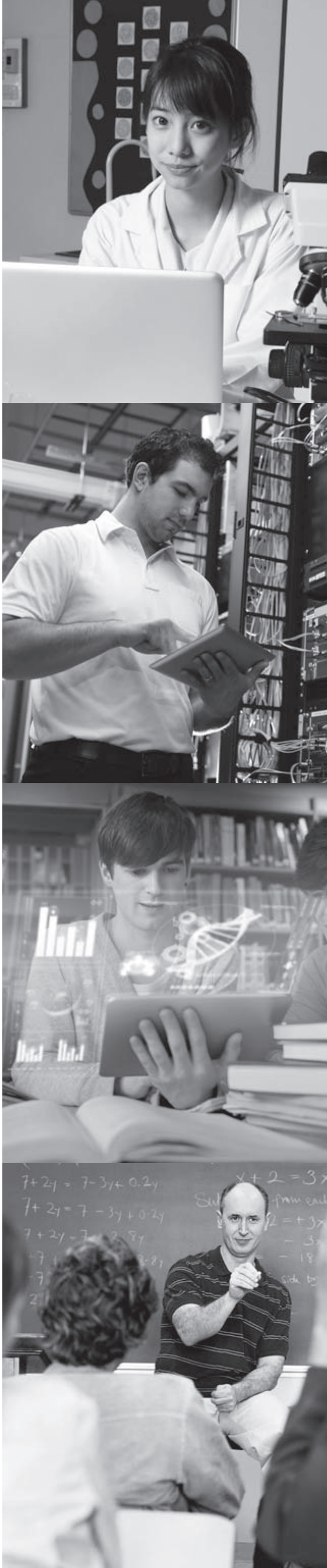
Purdue University Press, 2014.
Project MUSE.muse.jhu.edu/book/42542.

CHAPTER 6

TEACHING DATA INFORMATION LITERACY SKILLS IN A LIBRARY WORKSHOP SETTING

A Case Study in Agricultural and Biological Engineering

Marianne Bracke, Purdue University
Michael Fosmire, Purdue University



INTRODUCTION

This Data Information Literacy (DIL) project team worked with two faculty members in a hydrology lab in the Department of Agricultural and Biological Engineering at Purdue University; this was one of two Purdue University teams participating in the DIL project. The data produced by the lab include field-based observations, remote sensing, and hydrology models to help understand land-atmosphere interactions and the hydrologic cycle. Interviews with the faculty and graduate students in the research group indicated that data management standards were their primary concern. These Purdue researchers were neither aware of nor using disciplinary-developed data standards for storage, sharing, reuse, or description of data. Data standards would allow their data to be interoperable with other data generated by researchers in their field and would prevent them from "reinventing the wheel" each time data must be shared. Additionally, they were very interested in contributing to disciplinary standards since they believed that standards developed by the community had a better chance of being adopted. Over the course of the project, one of the participants became the campus representative to a national data repository, which gave our program a greater urgency: current and future students who worked in their labs must be trained in and use these standards.

Through user assessment, the DIL team members determined that the most important DIL areas to address through instruction were creating standard operating procedure documents for collecting the lab's data, finding external data, and creating metadata. With regard to operating procedures, the research group indicated that they had some instructions for data management listed on their wiki, but students did not follow them very often. The DIL team determined that the students had not internalized the need to manage and document data for their own work and to share with other members of the group. The wiki procedures were not specific enough to give students direction to successfully manage their data. Students also needed to incorporate external data—for example, using weather/climate data as inputs in their simulations. Locating, understanding, cleaning, and formatting those data is not a trivial process, and students can save significant time if the data are in a format that is usable by or easily importable into their programs. Finally, metadata was the key to effectively organizing, managing, and disseminating data. The more one knows about the contents of a data set, the more likely one can make the right choice about whether to use it. So, a well-documented data set will be more visible, comprehensible, and potentially useful to the research community at large.

We determined that the most effective approach to teach these skills within the time constraints of the research group was to conduct three instruction sessions over 3 months during the lab's normally scheduled meetings. Embedding the instruction within the lab's meeting schedule emphasized (1) how important the data skills were to the faculty members, and (2) that there was an urgent need to embed community standards for data management and curation into everyday practice. Overall, this approach to instruction was to present a contextualized program, grounded in the actual activities and procedures of the group, to reinforce the practical need for DIL skills and attitudes and increase buy-in from the lab group members.

We developed a different assessment for each module, appropriate for the range of learning objectives. The results of the assessment revealed that applying the content presented to real-life research workflows is a real challenge for students. Even though they clearly understood the material presented—and even recognized its importance—students did not incorporate data management practices into their everyday workflow. Future plans include collaborating with the faculty and students to incorporate these skills into standard lab practices.

LITERATURE REVIEW AND ENVIRONMENTAL SCAN OF DATA MANAGEMENT BEST PRACTICES

The literature review focused primarily on water and hydrology disciplinary data management resources, though the interdisciplinary nature of the lab's work led us to include ecological and biological research resources as well. The literature showed that students had little experience with creating metadata (Hernandez, Mayernik, Murphy-Mariscal, & Allen, 2012).

The most useful information for our background review came from the Consortium of Universities for the Advancement of Hydrological Science, Inc. (CUAHSI) organization (http://www.cuahsi.org/). Created in 2001 by the National Science Foundation, CUAHSI is the water-science community response to "the need to organize and extend the national and international research portfolio, particularly to develop shared infrastructure for investigating the behavior and effects of water in large and complex environmental systems" (CUAHSI, 2010). The consortium lists a number of points in its mission statement that are crucial to addressing better access to data, including creating and supporting research infrastructure and increasing access to data and information. Its strategic plan lists four data access goals, which demonstrate the forward thinking of the organization:

1. Develop and maintain search services for diverse sources of data and the underlying metadata catalogs (building on and extending from the Hydrologic Information System—HIS), including an access portal and coordination with providers of water-related information
2. Develop a mechanism for citation and use tracking to provide professional recognition for contributions to community data archives
3. Solicit community input on emerging data needs and facilitate access to new types of data
4. Coordinate development, promotion, and adoption of metadata standards between universities, governmental agencies, and the private sector for interpreted data products (e.g., potentiometric surfaces, areal estimation of precipitation, and input-output budgets). (CUAHSI, 2010, p.18)

Perhaps the most interesting area to note in the CUAHSI strategic plan is its continued development of metadata standards. CUAHSI recognizes the need for a shared language for both researchers and information systems to communicate to other researchers and information systems. To this end, the consortium is expanding the CUAHSI Hydrologic Information System (HIS), a Web-based portal for accessing and sharing water data (CUAHSI, 2013). The HIS operates with two important metadata standards: the Water Metadata Language (OGC, 2013), which is an open metadata schema created by the San Diego Supercomputing Center for hydrological time series and synoptic data, and the Federal Geographic Data Commission (FGDC) metadata schema (FGDC, 1998) created for geographic information system (GIS) and spatial data. Other metadata and data practices include the well-developed schema of the Ecological Metadata Language (EML), originally developed by the Ecological Society of America for ecology and related disciplines (Knowledge Network for Biocomplexity, n.d.b). Although not specifically created for hydrology, the EML metadata standard uses similar descriptions and requires an understanding of geospatial needs that are specific to the hydrology discipline, more so than more general standards such as Dublin Core (Dublin Core Metadata Initiative, 2013). Additionally, this Purdue DIL team consulted very useful EML tools, such as the Morpho data management application, a downloadable metadata entry template (Knowledge Network for Biocomplexity, n.d.a), when creating a metadata exercise for the graduate students.

Since the greatest needs for our research group focused on metadata and laboratory standard operating procedures for data management, we consulted Qin and D'Ignazio (2010), who provided details of a metadata-focused scientific data course of study. Stanton (2011) described the duties of practicing e-science professionals, which provided a foundation in actual tasks that scientists undertook in the course of managing data. Finally, the EPA (2007) provided a solid introduction to the purpose and process of creating standard operating procedures, which were applied to the student activities.

CASE STUDY OF GRADUATE STUDENT DATA INFORMATION LITERACY NEEDS IN AGRICULTURAL AND BIOLOGICAL SCIENCES

The hydrology research groups consisted of two faculty members who focused on the integration of field-based observations, remote sensing, and hydrology models to increase understanding of land-atmosphere interactions and the hydrologic cycle. Their work requires the acquisition of different kinds of data and the ability to convert data to ensure interoperability. The primary faculty member understood the importance and significance of good data practices, but still struggled with achieving high-quality data management in the research groups. The data collected in the lab ran the gamut of data types. On the one hand, the lab manually collected water samples and analyzed the results; tracking their processes with print lab notebooks that were later scanned into electronic formats. On the other hand, the group also downloaded remote sensing data from external sources, which were fed into computer models that created large data files in the process. Managing these three types of data—field samples, (external) remote sensing data, and computer simulations—provided constant challenges, especially as the students gathering or processing each different kind of data communicated their results with each other.

To understand the needs of the graduate students, the Purdue DIL team conducted six interviews between April and June of 2012. We used the DIL interview protocol (available for download at http://dx.doi.org/10.5703/1288284315510). This is a semi-structured interview instrument that allows for follow-up and clarification questions. The Purdue DIL team interviewed the primary faculty member (Faculty A), from the Department of Agricultural and Biological Engineering (ABE). We then interviewed five ABE graduate students (a mix of master's and PhD students) working in this faculty member's research group. (Note: A second faculty member [Faculty B] and other graduate students working on their research team could not be reached for interviews but were included in the educational program. This second faculty member was included in all subsequent actions and discussions in creating instructional content and assessments.)

One reason that our team approached Faculty A to be part of this project was because he had already expressed concern about teaching data management and data literacy skills to graduate students for the educating, acculturation, and training process of graduate school. He was familiar with many data literacy skills already, generally from the absence of good practices. These resulted in data loss by students due to the lack of proper backup, poor description, and poor organization of files. For example, he described:

> I have been slowly developing a data management plan after our conversations over the last couple of years, . . . [but one] that's more in my head. . . . But I think just the general conversation has clarified in my head that rather than just repeating over and over again to my students what they should be doing, having a written statement certainly helps. And then when they get in trouble, like the student who was saving everything on their external USB hard drive, I [can] point back to the data management plan that says [they] weren't allowed to do that.

He further described:

> I tried to establish a naming convention, but nobody ever listens to the naming conventions, so next thing you know you've got five files labeled "Final 1", "Final 2", "Final A", "Final C." So we keep running into this problem with stuff that people who have left, right? So what is this file? We've got three files that look identical except for the "Final" variation name. Which one is it?

Faculty A also experienced difficulties with understanding or obtaining the lab's data from students after their graduation. He explained:

> I had a student in my first couple of years who [collected] field data for me, and I didn't have a written plan. He didn't follow my [verbal] plan, and so he left with all of the material. . . . I've had a couple of people ask me about that data and what was available and it's like, well, I've never actually seen it.

Faculty A offers a class on environmental informatics. Most of the skills in the course are not taught to graduate students generally prior to their entering the lab unless they are picked up informally from other advisors or students. The class included general best practices for research, but many discipline-specific items were covered as well. Even so, one of Faculty A's primary concerns was that students were not receiving any data training outside of his lab or in their course work. Additionally, all his research group students were in the ABE department studying some aspect of hydrology but from a variety of angles: using field or observed data, using remote sensing data, or creating models. This meant that it was difficult to create and enforce a one-size-fits-all approach to a written DMP. Faculty A stated:

> So I think if you have a lab-based kind of group, then they probably have some methodology that they lay out in a lab book, but it's harder when it's—you know—a small group and people are doing different things. This is the dilemma for me. I've got one graduate student who's doing mostly remote sensing work. I've got a couple of grad students who are going to do more observational work. And then most of them are doing modeling work. . . . [I]t becomes more individualized, right? It's harder to invest the time to come up with the documentation [for data management] because it's [for] one or two people. But

> the problem is that those one or two people become somebody else [grad students replacing current] or maybe multiple people at some point, right? So we need to be capturing this.

To help with this problem, Faculty A had introduced students to some general data management policies on a wiki site once they started in his lab. When interviewed, students all displayed some awareness that there were formal data management policies in place within the research group. However, they also all expressed varying degrees of compliance, sometimes because they were not sure they applied to their specific data situation. One graduate student said:

> Yes we have a wiki site. [The faculty advisor] lists all of the procedures that we need to follow. . . . (Laughs) But I think I do not follow that, because my data is too large and it's very difficult to ask Purdue to extend my space.

In addition to our interview results in the DIL project, our interview included ratings of the DIL competences. Here, both the faculty and the graduate students interviewed rated most of the DIL facets as important (see Figure 6.1). The highest rated concepts by the students were *discovery and acquisition, data processing and analysis,* and *data management and organization,* with *ethics and attribution, data visualization and representation,* and *metadata and data description* very highly rated as well.

A MULTI-SESSION INSTRUCTION APPROACH TO DATA INFORMATION LITERACY SKILLS

In developing our DIL program, we discussed with both of the faculty members the nature and extent of instruction needed by their students. The discussion centered on the highest priority skills needed by the students, which skills would best be facilitated by librarian partners, and which skills, if successfully learned, would have the greatest impact on the research group overall. We also discussed how much time would realistically be available for face-to-face instruction, so that we could make the best use of the research groups' time. With a total of 2 faculty members and 13 students, each with their own academic schedule, the faculty found it challenging to find dates and times for even an hour-long group meeting a week.

We settled on a three-part instructional strategy that included some prep work prior to the face-to-face session and homework for the students to complete following the session. Given the time constraints, the DIL team felt that we should concentrate on just the most important and directly applicable DIL skills for which the librarians had unique expertise. Consequently, we decided to focus our instruction on *discovery and acquisition, data management and organization, ethics and attribution,* and *metadata and data description* as the remaining high-impact fundamental areas from the survey. While additional topics such as *data visualization and representation* and *data processing and analysis* were important, they might best be taught by the faculty members themselves.

It became apparent that, while the research group had a preliminary set of data management policies, these policies were not well understood or adhered to by the graduate students. Thus, we determined that one way to provide a scaffold for the DIL topics would be to develop standard practices for handling data in the research group. From the literature review and environmental scan, we concluded that these standards must be developed collaboratively to ensure maximum adoption by the group. In short, our goal was to help the group establish its own community standards.

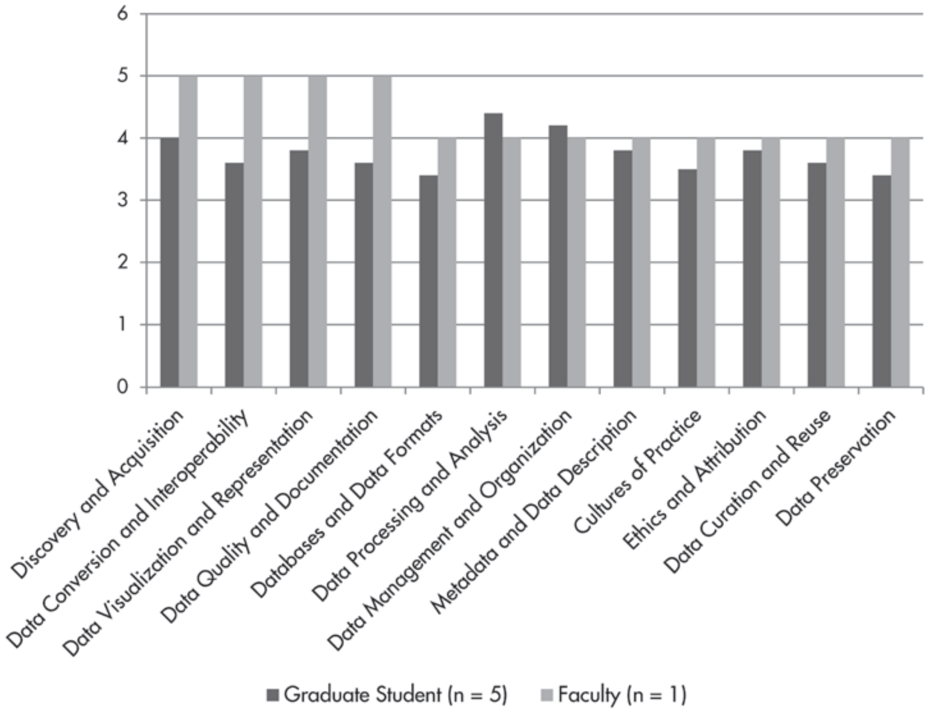


Figure 6.1 Average DIL competencies ratings for the agricultural and biological sciences case study. Ratings based on a 5-point Likert scale: 5 = essential; 4 = very important; 3 = important; 2 = somewhat important; 1 = not important.

To increase the authenticity of the exercises, each of the instructional activities focused on students tackling the actual problems of their group using the content presented in class.

RESULTS OF THE FALL 2012 INSTRUCTION SESSIONS

On the basis of our findings, our team decided to give three presentations to the combined research group over a 3- to 4-month period in the fall of 2012. Our approach was to fold the instruction into the regular meeting schedule to make the DIL material part of their workflow, rather than as something extra or outside of what they would have to do as a group anyway. Faculty A and Faculty B's research groups met together biweekly, so our team worked with them at every other meeting, or roughly once a month, starting in September, for a total of three sessions.

The topics for the three sessions included (1) developing a data checklist modeled on a standard operations procedures or laboratory protocol format, (2) searching for data in external databases, and (3) creating metadata. The learning objectives for each session are listed in Table 6.1 and the following sections detail the sessions.

Session 1: Data Checklist/Standard Operating Procedures

The aim of Session 1 was to teach the students to articulate the relevant components of a standard operating procedure and to apply those components when creating the actual procedures for the research group. In earlier

TABLE 6.1 *Learning Objectives of the Fall 2012 Library Instruction Sessions*

Session #	Topic	Learning Outcomes
Session 1	Data checklist/ standard operating procedures	Students are able to articulate the relevant components of a standard operating procedure and apply those components to create an actual procedure for the research group
Session 2	Searching for external data	Increased student appreciation for the value of metadata in locating data from external sources, and as a corollary, the importance of applying metadata to their own data sets so others can find (and cite) them in their own research
Session 3	Creating metadata	Students are able to analyze their own data sets and determine appropriate metadata to describe those sets. Students would then be able to curate their data within the structure of Purdue's data repository

discussions with Faculty A, he mentioned that something as simple and straightforward as a checklist for the kinds of data that might be collected would be a good approach. This could outline all the types of data needed, while providing an overview of the data in this outline. Faculty A created an initial checklist for the three categories of data collected: field observation data, remote sensing data, and model simulation data. Each category was unique and therefore had a different checklist governing its organization. Initially, each checklist contained 7 to 15 elements. For example, the field observation data checklist included the following information and data elements for organization and management:

- Field notebooks—scanned copies of all pages related to activities
- Digitized notes and measurements from field notebooks
- Raw files downloaded from field equipment
- Changes to sample control program (text file)
- Photos of sample sites
- IDs associated with physical samples, if collected
- Lab analysis results for all physical samples

The original checklist was meant to be a step-by-step list of things that a student might do to properly capture and describe all the data gathered in an instance of field observation. However, after discussions with the faculty collaborators, we determined that the checklists gave insufficient or ambiguous directions, which was why students did not find the checklists useful.

The DIL team started the session by having students recall when they started in the group and what information they would have liked to have about the data they were working with from the previous students. We brainstormed the attributes that were important to them (e.g., units, weather conditions, analysis techniques, calibration information) and used that to set the stage for determining how they could provide that information about the data they were collecting or producing. We also introduced some examples of best practices in standard operating procedures to show students how to translate their needs for information into an actual set of steps/activities that would lead to the production of that information.

The team followed up the instruction with an exercise using these checklists. To have the students gain ownership of the checklists, the team asked students which elements were missing. This generated some initial suggestions, and then we broke the students into three groups based on which of the three checklists matched most closely with the type of work they did within the research group. Some students matched with two or even all three areas, so they self-selected which group they wanted to join based on their interest or to help balance the group sizes. The faculty members each joined one of the groups. The groups were then asked to work with their assigned checklist in more depth, adding to it and documenting the most realistic way it could be implemented in current workflows. Their homework was to finish their checklist and share their work with the group in 2 weeks. Each group took a slightly different approach; the two groups with the professors as members were more thorough than the third group. The third group possibly lacked the pressure, the focus, and the expertise of having their instructor as a member of their work group.

The three resulting checklists are in Appendix A to this chapter, and the entire research group continues to work toward incorporating the data checklists into their regular workflow. Overall, the team found that the final, community-driven checklists were greatly improved over the faculty member's original draft. They exhibited more detail and less ambiguity, and they showed that students could transfer the content of the instructional session to documentation that was directly relevant to their lab.

Session 2: Searching for External Data

For the second session, the goal was to increase student appreciation for the value of metadata in locating data from external sources, and as a corollary, the importance of applying metadata to their own data sets so that others can find (and cite) them. After debriefing the checklist homework from the first session, which provided reinforcement of the core concepts of standard operating procedures, the second class introduced the Ecological Metadata Language or EML, and Morpho, the tool for describing data sets using EML. Although the Water Metadata Language (WML) at first seemed to be the best fit with the hydrology group, and may prove to be in the long run, the WML tools were not yet as fully developed nor as user-friendly as those provided for EML. The DIL team began the discussion with the "peanut butter sandwich exercise" (i.e., to write down the instructions to make a peanut butter sandwich and then have someone else carry out those instructions explicitly). This demonstrated how description can make a difference in how well individuals understand procedural processes and to illustrate the need to be explicit and complete when describing something.

Next, we drew parallels of the description exercise to metadata. Here we discussed how well-documented metadata could help someone else understand a data set—from how it was gathered to how it was analyzed—and its greater meaning in the context of other data. Students were divided into small groups and asked to search the Knowledge Network for Biocomplexity (KNB) data registry using Morpho to find a data set that might be relevant to them. This was challenging for many students: the keywords that they used were very specific and often unsuccessful while very general keywords such as "water" succeeded. The general "water" records were quite illustrative of how helpful more precise and in-depth descriptions would have been for the searcher.

In the end-of-class assessment, we asked students what they learned, what they will begin to incorporate into their own work, and what was still unclear (see Appendix B to this chapter for the assessment tool). Almost all students responded that they had a deeper understanding of how important metadata could be in describing their data to others and as a way for others to locate their data. They also appreciated the need to be explicit in their own descriptions of their data so that searchers can determine if and how the data might be useful to them. The results of these self-assessments, reinforced by the instructors' observations of the students while searching for external data, aligned very well with the learning outcomes. The students saw clearly that poor description could make another researcher's data difficult, if not impossible, to reuse, and this set the stage for what they would learn in Session 3, creating their own metadata.

Session 3: Creating Metadata

We designed the third session for students to be able to analyze their own data sets and determine appropriate metadata to describe those sets within the structure of an online repository. To demonstrate this, students were asked to submit their own data to our institutional data repository, the Purdue University Research Repository (PURR), and to create a brief metadata record to describe it. We asked students to bring a sample of their data to this session. A data scientist introduced the students to PURR and described the basic principles of what a repository could do for their submitted data. After a brief walk-through on the mechanics of getting started, which included creating an account in PURR, each student and the two faculty members created a project space. The PURR project space allows users to designate individuals with various roles such as "collaborators" or "owners," and allows owners of the project space to provide access to the materials in their project space to selected individuals. Each participant then uploaded his or her data file to the project space.

For each file uploaded, PURR requires very basic metadata, based on the Dublin Core metadata standard (http://dublincore.org), for description. Because the metadata that is asked for by PURR is so general in nature, we decided to add a more sophisticated metadata assignment to the class that was discipline appropriate. For this assignment, the libraries' metadata librarian created a Web-based form based on EML (see Appendix C to this chapter) and asked students to fill out and include with their data submission to PURR. The 15-field metadata form included subject-based items such as geographic coordinates, temporal coverage, methods, and sampling units, as well as more general items like keywords, abstract, data owners, and data contacts. This information automatically populated an Excel file that could be repurposed as a supplementary document for the data deposited into PURR. Unfortunately, at the time PURR did not accommodate custom metadata fields as a part of its metadata registry. So the metadata had to be downloaded as a separate text file for a potential user of the data to take full advantage of the EML information provided by the author. The metadata, if properly qualified, could also be inserted into a bibliographic data repository, such as the KNB data registry, using their metadata software, Morpho. However, the students were not asked to take that extra step due to time constraints.

This exercise required students to think about how best to describe their data for anyone other than themselves. This required them to capture their tacit knowledge and internalized assumptions about a data set—knowledge that must also be passed along to another individual, even someone they may be working closely with, in order for them to understand the data. DIL team members reviewed the students' metadata submissions and offered suggestions for improvement. Although students were reluctant to do additional metadata entry when depositing their data, the convenience and straightforwardness of the online form improved students' willingness and confidence to complete this task successfully. In the future, as the use of WML continues to increase and as it becomes more robust, we recommend using an online metadata form with fields from WML, or a blend of EML and WML, if that would be appropriate, for a broader audience of data submitters.

Although students said that they understood the need for good descriptive metadata, they were not quick to fill out the metadata template that we provided. Students were prompted several times to complete the form, and 10 out of 12 finally submitted the form. When filling out the forms, students succeeded in writing descriptive methods, study extent, and sampling procedures, and to a lesser extent, in providing keywords (perhaps because completing these tasks are already a familiar exercise when writing papers for journals). Additionally, they were very thorough in describing geographic coverage. This may not be surprising given the geographic focus of their research. Students were less successful when listing data owners, contacts, and affiliated parties, even though this was covered in class. Understanding who owns the data and what roles they "officially" play in creating the data was a complicated aspect of describing data. This is an area that the team intends to cover more fully in future sessions. Overall, the team will need to find ways to work with the faculty members to insert the metadata template into an existing workflow, so that students do not see this merely as something externally imposed and extra work.

DISCUSSION

The integrated lab-meeting approach was generally successful and contained elements that could be replicable for a wider audience. The exercise of creating checklists to address *data management and organization* skills, though the results here are specific for these research groups, is a general approach that could be used by other labs or researchers. Any lab or work group can generate the detailed list of items that need to be captured or addressed in the data gathering process. Also, with the faculty-student-librarian team approach used in the DIL project, this list can be developed so that there is a feeling of shared ownership and responsibility, each bringing unique skills and responsibilities to the task. Faculty provide the domain expertise and an understanding of what information absolutely has to be collected. Students bring an operational perspective of how the data are incorporated into the data collection; they are often the ones performing the collection tasks and can identify ways to streamline the process. Finally, librarians bring the DIL expertise to facilitate the discussion between faculty

Contextualize the DIL model to the needs of the target audience and highlight specific benefits of data management skills for each research group.

and students as well as to optimize the accessibility, internal consistency, and organization of the data.

Even before the DIL project began, the disciplinary faculty member believed that metadata, or some description of the data, was critical. He had experienced too many instances where one student's data could not be understood, by himself or by others, due to inadequate description. Sometimes this was reparable after many hours spent trying to reconstruct what the data represented; other times the data were simply lost or unusable due to the fact that the description could not be recovered or the student had graduated and taken the data. Our instruction sessions covering the importance of good data description and specific metadata tools positively impacted the students' understanding of the issue. In their assignments the students demonstrated their understanding of how poor metadata could make a data set useless to anyone other than the creator. They applied this knowledge when creating better metadata for their own data descriptions meant for a broader audience.

Despite this appreciation, the students still needed metadata tools to guide this process if they were to be successful. Creating the online tool for entering modified EML metadata increased the likelihood that they would actually adopt this new step in the data management process. The DIL team would like to make the metadata more usable, so that others might take advantage of the work that the students put into describing their data. Currently, saving the EML metadata as an Excel file does not take full advantage of the power of the descriptive language; therefore developing a more robust online entry form and/or brokering the metadata to disciplinary-specific repositories will help students appreciate the value of their work. Ultimately, search tools that take advantage of the descriptive metadata can lead to greater reuse of the data by others.

Getting the students to adopt these practices into their everyday workflow was a challenge.

However, getting the students to adopt these practices into their everyday workflow was a challenge, and we had limited success with this during the project. In hindsight, recognizing adoption as one of the greatest barriers, we might have worked with the students from the beginning to incorporate these practices into their research workflows. In tandem, we might have worked more closely with the faculty to create a structure, higher expectations, and a process for implementing the DMP within the lab. However, the adoption of these new practices might simply take time. It could be that regular use of the practices will eventually become habit. Additionally, asking the faculty partners to enforce the new practices through regular and frequent monitoring will likely pay off in the long run with regard to adoption. As these practices become "business as usual" they will transfer easily to new students as they cycle into the research groups and formal training for one student becomes peer-to-peer learning for the next.

CONCLUSION

Overall, this DIL team felt that the program was very successful in communicating DIL concepts and impressing upon graduate students the importance of good data practices. Implementation is still a work in progress, as the faculty researchers are in the best position to address accountability in order to embrace the practices that the group has developed. That said, there have been robust conversations within the research group about the need for improving data management, and all of the members of the group are speaking from a higher level of understanding than they had previous to the project. The DIL model works best when contextualized to the needs of the target audience. Hands-on activities aligned with the goals of the research group extended what they were already doing or trying to do, which gave them more tools and concepts to apply to their research environment. At the end of the instructional program, students had tangible results that included standard operating procedures for the lab and data sets submitted to a repository.

As we reflect on the activities, *data management and organization* (standard operating procedures) and *metadata and data description* (describing and depositing data sets into a repository) jump out as the areas that found the most traction within the research group, and might be the driving principles for a more general DIL model in this discipline. Also, while library and information science professionals may focus on the need to share data and make it openly available, the focus among researchers is shifted more toward sharing data and making it accessible mainly within the research group. Therefore, when stressing the value of data management skills, highlighting the benefit to the research group is key.

In the course of the activities, we discovered that much of the data in distributed repositories is not well described, so locating and using that data is a continuing challenge. As a result, researchers may gravitate toward centralized, well-stewarded data—for example, such as that produced by government agencies. For many "small science" areas, the lack of quality knowledge management systems provides challenges for the successful interoperability and sharing of data among research groups. The lack of good metadata limits progress in this area, as there are few examples of best practices in action in the disciplinary data repositories for their community.

Finally, this case study found that graduate students have no trouble grasping the concepts of DIL when the concepts are presented to them. However, getting students to change current practices, whether on their own or in a group setting, is an ongoing challenge. It is unclear whether this is due to the lack of emphasis on data management in the lab, because faculty are not stressing the need, or that students are not comfortable nor knowledgeable about how to adjust current practice. The important conclusion is that our educational approach of modules was not enough to ensure implementation of best practices. Further research and development is needed to address how students and faculty can not only learn the skills involved with DIL, but implement the DIL best practices as well.

ACKNOWLEDGMENTS

Thanks to C. C. Miller, who contributed to the initial stages of our Purdue DIL team project.

NOTE

This case study is available online at http://dx.doi.org.10.5703/1288284315478.

REFERENCES

CUAHSI. (2010). *Water in a dynamic planet: A five-year strategic plan for water science.* Retrieved from https://www.cuahsi.org/Posts/Entry/115292

CUAHSI. (2013). The CUAHSI hydrologic information system. Retrieved from http://his.cuahsi.org/

Dublin Core Metadata Initiative. (2013). The Dublin Core Metadata Initiative. Retrieved from http://dublincore.org/

Environmental Protection Agency. (2007). *Guidance for preparing standard operating procedures (SOPs).* Retrieved from http://www.epa.gov/quality/qs-docs/g6-final.pdf

Federal Geographic Data Committee. (1998). *Content standard for digital geospatial metadata workbook, Version 2.0.* Retrieved from http://www.fgdc.gov/metadata/documents/workbook_0501_bmk.pdf

Hernandez, R. R., Mayernik, M. S., Murphy-Mariscal, M., & Allen, M. F. (2012). Advanced technologies and data management practices in environmental science: Lessons from academia. *Bioscience, 62*(12), 1067–1076. http://dx.doi.org/10.1525/bio.2012.62.12.8

Knowledge Network for Biocomplexity. (n.d.a). Morpho data management application. Retrieved from https://knb.ecoinformatics.org/morphoportal.jsp

Knowledge Network for Biocomplexity. (n.d.b). Ecological metadata language (EML). Retrieved from https://knb.ecoinformatics.org/#external//emlparser/docs/index.html

OGC. (2013). OGC WaterML. Retrieved from http://www.opengeospatial.org/standards/waterml

Qin, J., & D'Ignazio, J. (2010). The central role of metadata in a science data literacy course. *Journal of Library Metadata, 10*(2–3), 188–204. http://dx.doi.org/10.1080/19386389.2010.506379

Stanton, J. M. (2011). Education for escience professionals. *Journal of Education for Library and Information Science,* (2), 79–94.

APPENDIX A: Data Archiving Checklists for Session 1 of the Agricultural and Biological Sciences Case Study

These checklists were generated by the students and faculty in the Agricultural and Biological Sciences case study of the DIL project. They include checklists for handling the three types of data generated by the research group: (1) Field Observation Data, (2) Remote Sensing Data, and (3) Simulation Model Data.

Data Archiving Checklist
Field Observation Data

Field notebooks—scanned copies of all pages related to activities
  Date scanned:
  Date scanned:
  Date scanned:
Digitized notes and measurements from field notebooks
  Date scanned:
  Date scanned:
  Date scanned:
Raw files downloaded from field equipment
  Date downloaded:
  Date downloaded:
  Date downloaded:
Changes to sample control program (text file)
  Text file name:
  Photos of sample sites
  Photo files stored:
IDs associated with physical samples, if collected
  ID:
  ID:
  ID:
Lab analysis results for all physical samples
  Files stored:
  Files stored:
  Files stored:
Associated remote sensing data?
  Notes:
Associated simulation data?
  Notes:

Processed Files

Name of file:
Quality control program:
Outside data sources:
Photographs of samples:
Writing/compiling data:
Order of Processing:
  Formats, fields, missing data, processing, units, time, how collected, where collected, weather conditions
  Simulation data: inputs, sim software used (version), size/resolution/scale, format, fields, units
  Remote sensing: resolution-temporal, spatial; when collected; name of sensor; cloud/weather, calibration, projection, file type—raster/shape
Metadata file: Data dictionary

Data Archiving Checklist
Remote Sensing Data

Remote sensing platform(s) and sensor(s) used, and status
  Platform/sensor/status:
  Platform/sensor/status:
  Platform/sensor/status:
Raw remote sensing files (DNs)
  DN file:
  DN file:
  DN file:
Atmospheric conditions, including radiosonde or other vertical profile data; output from data assimilation models; weather maps—collect all available data
  Notes:
All files/information required to georegister imagery
  Files stored:
  Files stored:
  Files stored:
  Radiance files, not georegistered
  Files stored:
  Files stored:
  Radiance files, georegistered
  Files stored:
  Files stored:
Final imagery analysis products
  Files stored:
  Files stored:

- Documentation of all steps taken in processing remote sensing images to final form
  - Atmospheric corrections
  - Emissivity corrections
  - Georegistration process
  - Classification or analysis methods
- Associated field observation data?
  - Notes:
  - Associated simulation data?
  - Notes:
- Data dictionary

Data Archiving Checklist
Model Simulation Data

- Model inputs (all inputs should be for simulations used in analysis)
  - Meteorology
    - File stored:
  - Vegetation
    - File stored:
  - Soils
    - File stored:
  - Global control file
    - File stored:
  - Streamflow routing model input files
    - File stored:
    - File stored:
    - File stored:
- Model evaluation data
  - Observed streamflow
    - File stored:
  - Other observation types
    - Observation type/file stored:
    - Observation type/file stored:
- Model version
  - Hydrology model source code as used in the simulations
    - Source code file stored:
  - Routing model source code
    - Source code file stored:
  - Source code from other models used
    - Source code file stored:
    - Source code file stored:
  - Model analysis products

- Raw model simulation output
  - For very large data sets, a filename should be provided and a location on fortress
    - File stored:
  - For smaller data sets, all output files should be migrated into HDF5 or tarred into a single file
    - File stored:
- Files that have been developed from the raw model output and that were the basis of analysis (e.g., output from the HDF5 summary statistics program), especially if they contain additional information not used in the final published product but could be used for additional analysis
  - File stored:
  - File stored:
- All data files used to develop graphics or tabular data
  - File stored:
  - File stored:
  - File stored:
- Scripts used to develop published graphics or tabular data
  - Script:
  - Script:
  - Script:
- High-quality EPS (preferred since they can be edited for minor changes), PNG (figures), or JPEG (pictures) files of published figures.
  - EPS file:
  - PNG file:
  - JPEG:

Associated field observation data

- Notes:

Associated remote sensing data?

- Notes:

What not to do format/units

Metadata document/data dictionary

APPENDIX B: Assessment Tool for Session 2 of the Agricultural and Biological Sciences Case Study

The Data Information Literacy (DIL) team used the following tool to assess the students' response to each of our three sessions.

1. Briefly describe what you learned in today's session:
2. List one thing that you will definitely incorporate into your own data gathering/description/management after today:
3. Briefly describe anything that was discussed today that is still unclear for you:

APPENDIX C: Metadata Form for Session 3—Data Package Metadata

Enter the title of the data package. The title field provides a description of the data that is long enough to differentiate it from other similar data.

Title:*

Enter an abstract that describes the data package. The abstract is a paragraph or more that describes the particular data that are being documented. You may want to describe the objectives, key aspects, design, or methods of the study.

Abstract:*

Enter the keywords. A data package may have multiple keywords associated with it to enable easy searching and categorization. In addition, one or more keywords may be associated with a keyword thesaurus, taxonomy, ontology, or controlled vocabulary, which allows the association of a data package with an authoritative description definition. Authoritative keywords may also be used for internal categorization. An example of an authoritative thesaurus is the National Agricultural Library Thesaurus: http://agclass.nal.usda.gov/dne/search.shtml

Authoritative keyword source. If an authority was used for the keywords, identify by name the authority source.

Keywords (separate with commas):*

Enter information about the owners of the data. This is information about the persons or organizations certified as data owners (e.g., principal investigator for a project). The list of data owners should include all people and organizations who should be cited for the data. Minimally include full name, organization name, owner address, and e-mail.

Data Owners:*

Enter information about the contacts. This is information about the people or organizations that should be contacted with questions about the use or interpretation of your data package. Minimally include full name, organization name, contact address, and e-mail.

Contacts:*

Enter associated parties' information. These are persons or organizations functionally associated with the data set. Enter the relationship. For example, the person who maintains the data has an associated function of "custodian." Minimally include functional role, full name, organization name, party address, and e-mail.

Associated Party:

Is your data set part of a larger umbrella project? Data may be collected as part of a larger research program with many subprojects, or they may be associated with a single, independent investigation. For example, a large NFS grant may provide funds for several primary investigators to collect data at various locations.

If part of a larger project, identify by name the project. If applicable, include funding agency and project ID.

Enter a paragraph that describes the intended usage rights of the data package. Specifically, include any restrictions (scientific, technical, ethical) to sharing the data set with the public scientific domain.

Usage rights:*

Enter a description of the geographic coverage. Enter a general description of the geographic coverage in which the data were collected. This can be a simple name (e.g., West Lafayette, Indiana) or a fuller description.

Geographic coverage:*

Set the geographic coordinate s which bound the cove rage or a single point. Latitude and longitude values are used to create a "bounding box" containing the region of interest (e.g., degrees/minutes/seconds N/S/E/W) or a single point.

Bounding box or point:

Enter information about temporal coverage. Temporal coverage can be specified as a single point in time, multiple points in time, or a range thereof.

Temporal Coverage:*

Enter method step description. Method steps describe a single step in the implementation of a methodology for an experiment. Include method title, method description, and instrumentation.

Methods:

Study extent description. Describe the temporal, spatial, and taxonomic extent of the study. This information supplements the coverage information you may have provided.

Study extent:

Sampling description. Describe the sampling design of the study. For example, you might describe the way in which treatments were assigned to sampling units.

Sampling:

*Required fields