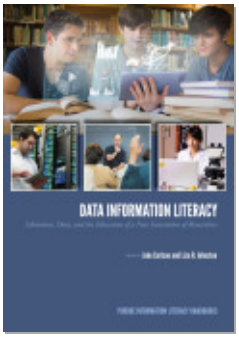
Data Information Literacy

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 .

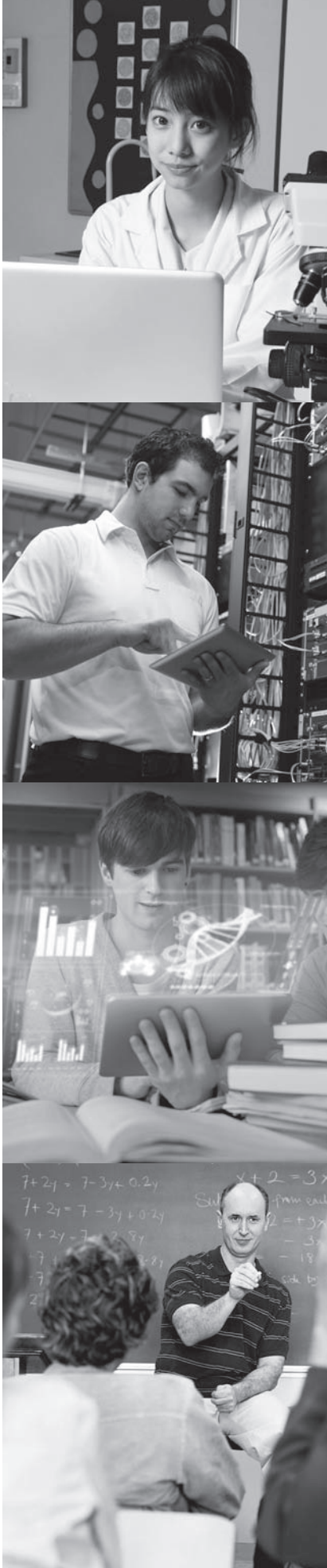
Purdue University Press, 2014.
Project MUSE.muse.jhu.edu/book/42542.

CHAPTER 3

AN EXPLORATION OF THE DATA INFORMATION LITERACY COMPETENCIES

Findings From the Project Interviews

Jake Carlson, University of Michigan
Jon Jeffryes, University of Minnesota
Lisa R. Johnston, University of Minnesota
Mason Nichols, Purdue University
Brian Westra, University of Oregon
Sarah J. Wright, Cornell University



INTRODUCTION

This chapter delves into the results of the user needs assessments we conducted for the Data Information Literacy (DIL) project and introduces the instructional interventions we developed to address those needs. Between March 2012 and June 2012, the five DIL project teams collectively interviewed 25 researchers (8 faculty and 17 graduate students or postdocs) on their DIL (instrument available at http://dx.doi.org/10.5703/1288284315510). We begin this chapter by presenting the broad themes that were uncovered across the interviews from our analysis. We then turn our attention to the responses given to each of the 12 DIL competencies by the faculty and students that we interviewed.

RESULTS OF THE DATA INFORMATION LITERACY INTERVIEWS

The results of the five case studies (presented in Chapters 4 through 8) revealed similarities and differences between faculty and students in how they perceived the importance of the DIL competencies for graduate students. Due to the small sample size and the use of convenience sampling, these results cannot be generalized outside of these case studies as indicators of each disciplines' importance ranking. Nevertheless the findings offer a useful starting point for larger investigations into the current environment of the educational needs of graduate students.

The DIL competency ratings based on a 5-point Likert scale are displayed in Figure 3.1. They show that, on average, participants valued each competency as either "important," "very important," or "essential." However, there was considerable variance in the responses received as indicated by the high standard deviations (ranging from .75 to 1.02). The competencies that pertained more directly to keeping a research lab operational and to publishing outputs, such as *data processing and analysis, data visualization and representation,* and *data management and organization,* tended to be rated more important than competencies that are less central to these activities, such as *discovery and acquisition* and *data preservation.* Although deemed important, some of the lower rated competencies, such as data preservation, are difficult to address. In the interviews, many faculty stated that they lacked the experience or knowledge to educate students effectively about these competencies. Several of the faculty and students questioned whether their field had a culture of practice in managing, handling, or curating data.

Figure 3.1 also shows the differences in how the participants viewed some of the competencies. Faculty generally placed a higher value on student development of competencies in actively working with data (e.g., *data processing and analysis, data visualization and representation*) and in competencies that would sustain the value of the data over time (e.g., *metadata and data description, data quality and documentation*) than the students did. Students gave the *discovery and acquisition of data* competency a higher rating than did the faculty. Students indicated in the interviews that this was an important component of learning their field and contextualizing their research. Two of the faculty who worked with code as their data gave *data management and organization* a lower rating than did the other participating faculty. One faculty member believed that, individually, students should know how to manage their own data but did not necessarily need to know

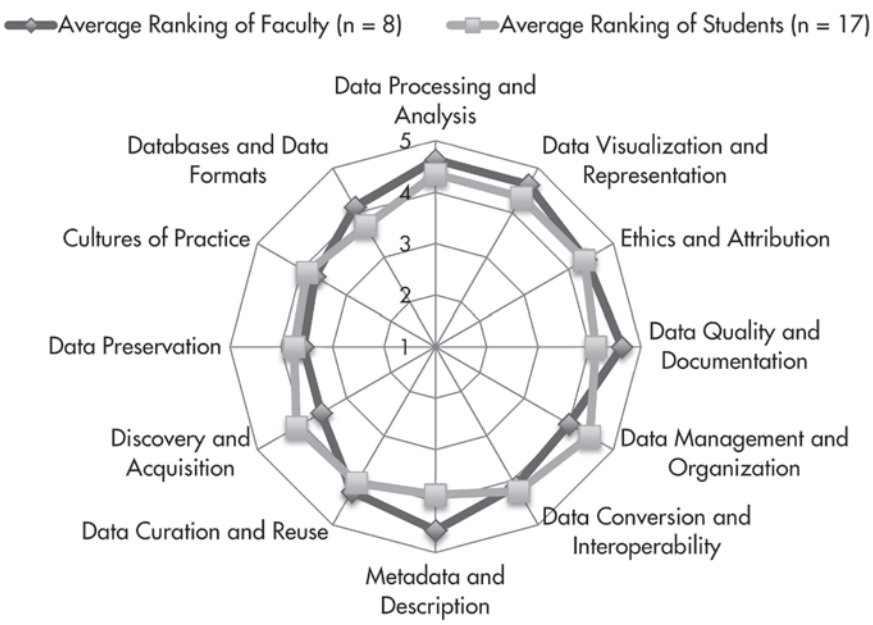


Figure 3.1 Graphical comparison of faculty and student ratings of importance of DIL competencies. Scale: 5 = essential; 4 = very important; 3 = important; 2 = somewhat important; 1 = not important.

how to develop systems or management plans for larger units. The other found it difficult to respond, not knowing what constituted good management practice and therefore unable to say if it would be worth the investment of his and the students' time.

THEMES FROM THE DATA INFORMATION LITERACY INTERVIEWS

Analyzing the interview transcripts revealed several commonalities across the five case studies: the lack of formal training in data management, the absence of formal policies governing lab data, self-directed learning through trial and error, and a focus on mechanics over concepts.

None of the five research groups provided their students formal training in data management. Instead, faculty reported that they expected that their students had acquired most of these and other competencies prior to joining their lab. As a University of Oregon faculty member noted, "[students may have] picked up [their skills] at on-the-job training, because a lot of them had a former life in a professional field . . . or [it's] something they got as an undergraduate." In contrast, student interviews revealed wide variations in their prior experiences with data. Most of the students had attended a seminar on responsible conduct of research (research ethics) but reported that data practices were not covered in the seminar. Moreover, these students could not recall the specifics of what was stated about data practices. It should be noted here that none of the five case studies involved data that would require training on dealing with human subjects or sensitive data.

In lieu of formal training, most graduate students learned data management through

trial and error, reading manuals, asking their peers for help, or searching the Internet. Of the five labs participating in this project, only one had written policies for the treatment and handling of data. Respondents predominantly expressed disciplinary norms and processes for data management as underlying expectations that tended to be delivered informally and verbally. Some of the students interviewed had inherited data from previous students or others in the lab; this transference process also tended to be informal with minimal introduction to the data.

Faculty expected their graduate students to be independent learners. For example, one faculty member summed up the skills acquisition process as the "pain and suffering method," which she described as "[graduate students] try it, they fail, they see what failed, they come back to their advisor and you say, 'Ah, well maybe you should try X.' It is not something that we have attempted to teach, certainly."

When asked how well their students had mastered the DIL competencies, faculty stated that students tended to focus more on the mechanics of working with or analyzing data rather than the theories and assumptions underlying the software or tools they used. For this reason, some of the faculty expressed concern that students' understanding of these competencies may be somewhat superficial. For instance, one faculty member stated that students may be able to collect data from a sensor, but they did not necessarily understand the equipment variables that might impact data quality or accuracy. They may be more focused on getting the data than on understanding the steps and settings that created it. Similarly, some faculty felt that though students may be able to use tools to work with data, they did not always use them very effectively or efficiently. For example, one faculty member commented, "I certainly think that they learn basic visualization tools, but there's a difference between learning how to draw a histogram and how to draw a histogram that's informative and easy to read."

This differentiation between basic project-driven skills and deeper, transferable understanding is found in questions about managing and curating data. Most students described idiosyncratic methods of data management, and generally overestimated the capacity of their methods to support local collaboration. Only 3 of the 7 faculty interviewed felt that their students provided enough information about their data for the faculty member to understand it. Only one faculty member thought that students provided enough information for a researcher outside of the lab to understand and use the data. In contrast, 15 of the 17 students believed that they provided sufficient information for someone outside of the lab to understand and use the data.

Faculty wanted their students to acquire a richer understanding and appreciation for good data management practices, but there were several barriers that restricted faculty from taking action. First, spending time on data management was not a priority if it distracted from or delayed the research process. Faced with this pressure, faculty accepted that a minimal skill set was sufficient for their students to succeed in school. One faculty member stated, "[Students] can do their work without understanding this. It's not essential that they have this. It's best if they do, but they don't. I guess I could be doing more, but we don't talk about all of these functions. . . . I'm not sure they all understand why data has to be curated."

Second, faculty did not see themselves as having the knowledge or resources to impart these skills to their students themselves. One faculty member mentioned requirements by

funding agencies for data management plans and journals accepting supplemental data files as positive steps, but researchers in her field were ill-prepared to respond. Most of the faculty stated that there were no best practices in data management in their particular field. Faculty in this study did not believe that funding agencies, publishers, or scholarly societies in their discipline provide the guidance or resources to support effective practices in managing, sharing, or curating data. In the absence of such support, the data practices in their labs remain centered on local needs.

It is interesting to note similarities between our findings and the findings of others who have studied faculty perceptions of student competencies in information literacy. Shelley Gullikson (2006) surveyed faculty at institutions in eastern Canada to understand their perceptions of the ACRL Information Literacy Competency Standards. Her results indicated a consensus that information literacy competencies were important overall, but little agreement on when they should be taught. Claire McGuiness (2006) conducted semi-structured interviews with sociology and civil engineering faculty in the Republic of Ireland and found that faculty believed that students were acquiring information literacy competencies without formal or direct instruction but through other existing learning situations and course work. More recently, Sharon Weiner (2014) surveyed faculty at Purdue University to develop an understanding of to what extent information literacy concepts were taught by faculty across the disciplines. In addition to revealing significant differences between what aspects of information literacy were taught between the schools, faculty responses indicated that they expected their students to know how to avoid plagiarism, search for information, and define a research topic before enrolling in their courses.

FINDINGS ON EACH OF THE 12 DATA INFORMATION LITERACY COMPETENCIES

The rest of this chapter will discuss findings on the 12 DIL competencies across the interviews conducted by the five DIL project teams. Subsequent chapters describe the more specific findings by each project team and how the teams translated these findings into educational programs. Each of the competencies presented here includes the loosely worded skills description that was provided to the interviewees to ground the discussion, as well as any additional skills that they themselves articulated. Next, we summarize a curated list of responses from both faculty and students.

Cultures of Practice

Table 3.1 summarizes the results of our interviewee responses regarding the *cultures of practice* competency.

Faculty Responses

A major concern of faculty was the amount of prior training graduate students received with respect to cultures of practice for data. One faculty member described students' knowledge in this area as "underwhelming." Faculty felt that though students adequately saved their files and made backup copies, they were not as competent with sharing, curating, and preserving data. On the other hand, several faculty members commented that they themselves were unaware of any established practices, values, or norms for a data "culture of practice" in their discipline. For example, a computer science faculty member pointed out that knowing how to document research properly, and being able to go back to it in the future, is a discipline-wide issue.

TABLE 3.1 *Faculty and Student DIL Competency Ratings of Importance: Cultures of Practice*

Competency-related skills:	Recognizes the practices, values, and norms of chosen field, discipline, or subdiscipline as they relate to managing, sharing, curating, and preserving data Recognizes relevant data standards of field (e.g., metadata, quality, formatting) and understands how these standards are applied
Additional skills:	Identifies standard protocols in the lab that may or may not match discipline-wide standards
Faculty and student ratings:*	Faculty average = 3.71 Student average = 3.88

*Ratings based on a 5-point Likert scale: 5 = essential; 4 = very important; 3 = important; 2 = somewhat important; 1 = not important.

Overall, faculty believed that guidance in this area would be beneficial. While it's true that faculty recognized the importance of obtaining skills through experience or peer teaching, they would like to have formal training available so that established practices and norms might be followed in the lab and the discipline. One participant described an ideal course for learning cultures of practice in the discipline that would include attitudes, shared skills (e.g., scripting language), visualization techniques, and technical writing training for describing results according to cultural norms.

Student Responses

The students we interviewed were unaware of any standards or discipline-wide norms for organizing, documenting, and sharing data. Yet, they recognized that this would be useful and important. One student stated that if researchers did not adhere to the standards of their field, "the results will not mean as much." And several students mentioned that they would follow standards if such standards exist. One computer science student mentioned that metadata standards in academia and industry appear to be at odds, with a greater amount of metadata being required in industry. As many graduate students take positions outside of academia after graduation, developing an understanding of industry norms and expectations in working with data is a critical element of effective educational programs.

Data Conversion and Interoperability

Table 3.2 summarizes the results of our interviewee responses regarding the *data conversion and interoperability* competency.

Faculty Responses

Most faculty reported that competencies with *data conversion and interoperability* were generally underdeveloped in students. Faculty reported that their students acquired their knowledge and skills in this competency through classes, peers, and experience. One faculty member stated that his students needed more experience with how conversion can affect their data. Another mentioned that students need to be aware of issues surrounding data loss during data migration and have an understanding of appropriate open standards for file formats.

Potential data loss in the conversion process was mentioned repeatedly. Faculty reported

TABLE 3.2 *Faculty and Student DIL Competency Ratings of Importance: Data Conversion and Interoperability*

Competency-related skills:	Is proficient in migrating data from one format to another Understands the risks and potential loss or corruption of information caused by changing data formats Understands the benefits of making data available in standard formats to facilitate downstream use
Additional skills:	Understands the advantages of different file formats Ability to code
Faculty and student ratings:*	Faculty average = 4.13 Student average = 4.24

*Ratings based on a 5-point Likert scale: 5 = essential; 4 = very important; 3 = important; 2 = somewhat important; 1 = not important.

that students were not considering the potential for loss or corruption when converting their data files. One faculty member made a connection between understanding how data can be manipulated and ensuring the quality of the data. Another saw this as an important skill for students to develop not just for working in his lab but also for gaining employment after graduation.

Student Responses

Nearly all of the students (14 out of 17) reported converting data as a part of their work in the lab, though most did not mention conversion as a distinct stage of the data life cycle. Students responded to questions of data conversion and interoperability by discussing conversion techniques for raw data (i.e., Microsoft Access files to plain text files; proprietary sensor data to Microsoft Excel) as well as processed data (i.e., converting images created in gnuplot to GIF or JPEG; converting a figure to a table). Conversions ranged from a simple cut-and-paste transportation of data to identifying the meaningful elements of the data and extracting them into a usable format. Students were less concerned with data loss during the conversion process than faculty. A few students reported checking the data after converting them to ensure that data loss had not occurred.

Data Curation and Reuse

Table 3.3 summarizes the results of our interviewee responses regarding the *data curation and reuse* competency.

Faculty Responses

Faculty viewed *data curation and reuse* as an important subject, but commented that both students and the researchers themselves lacked these skills. In fact, several commented that the idea of data reuse is just beginning to take hold. One faculty member commented that the entire research lab needed a better understanding of who would benefit from data curation. Another felt that students generally don't have to concern themselves with these skills as the researcher decides when and how to make the data available for reuse.

Faculty also had a more personal reason for believing data curation and reuse to be important. In their experience, their data could not be recreated over the course of extended experiments and consequently must be curated. Therefore they were the number one

TABLE 3.3 *Faculty and Student DIL Competency Ratings of Importance: Data Curation and Reuse*

Competency-related skills:	Recognizes that data may have value beyond the original purpose, to validate research, or for use by othersIs able to distinguish which elements of a data set are likely to have future value for self and for others Understands that curating data is a complex, often costly endeavor that is nonetheless vital to community-driven e-research Recognizes that data must be prepared for its eventual curation at its creation and throughout its life cycle Articulates the planning and activities needed to enable data curation, both generally and within local practice Understands how to cite data as well as how to make data citable
Additional skills:	None
Faculty and student ratings:*	Faculty average = 4.25 Student average = 4.06

*Ratings based on a 5-point Likert scale: 5 = essential; 4 = very important; 3 = important; 2 = somewhat important; 1 = not important.

reuse consumers of their own data. Similarly, faculty commented that the academic culture places less emphasis on functionality of data for public use and rather focuses more on the researchers' needs. Not all data are viable for curation, however, as one faculty member noted; nonstandard code was not reusable and didn't promote future research.

Faculty were also asked whether they or their graduate students had ever deposited data into a data repository. Of the eight faculty interviewed, three had deposited data in a repository, three had not, and two did not answer the question. Those that had, deposited their code into SourceForge or Google Code. However, faculty reported that getting the software in a format in which it could be shared was difficult.

Student Responses

Students identified at which stages their data (raw vs. processed vs. published) would be most valuable to save, but the potential value for reuse in the data they created was not an immediate concern. Rather, students did not appear to understand the practices and skills that would be needed to support the reuse of their digital information. For example, one student believed that individuals in the lab were taking the necessary steps to prepare the generated data for eventual reuse, but was unsure of "exactly what they're doing."

Of the 18 students interviewed, 7 indicated that they had deposited data into a repository for reuse, though some of them indicated that these repositories were for a particular agency and not publicly accessible. Students were almost evenly split about their intent to deposit data into a repository in the future, with 7 indicating that they were planning to do so and 6 stating that they were not. Four students responded "I don't know" to the question. Almost all of the students we interviewed were willing to share their data with someone outside of their lab, with only one student responding "no" and one other stating "I don't know." Several students said they would need their advisor's approval before sharing their data. However, 12 of the 15 students who indicated they

TABLE 3.4 *Faculty and Student DIL Competency Ratings of Importance: Data Management and Organization*

Competency-related skills:	Understands the life cycle of data, develops data management plans, and keeps track of the relation of subsets or processed data to the original data sets Creates standard operating procedures for data management and documentation
Additional skills:	Familiarity with tools for data management Ability to annotate data sets at a higher level to keep track of changes and analyses performed
Faculty and student ratings:*	Faculty average = 4.00 Student average = 4.47

*Ratings based on a 5-point Likert scale: 5 = essential; 4 = very important; 3 = important; 2 = somewhat important; 1 = not important.

would share their data also stated that they would place conditions on sharing the data. The other 3 students responded "I don't know." The most common condition was that the student or the lab receives proper credit through a citation if the data were used in a publication. Other conditions mentioned were no redistribution of the data before publication of the findings of the lab of origin, and assurance that the data would not be misinterpreted by the recipient.

Data Management and Organization

Table 3.4 summarizes the results of our interviewee responses regarding the *data management and organization* competency.

Faculty Responses

Faculty described data management skills as standard operating procedures passed on from one student to the next. They believed that students gain rudimentary skills in data management in statistics courses prior to their graduate school career. "Learning by doing" was cited by many faculty as how students obtained these skills. If students were not proficient in this area, several problems arose, including code overwrites, haphazard organization, and the inability to locate specific data. Faculty also cited participation in internships as a way that students obtained proficiency.

Data management plans ranked as very important; however, faculty clarified that students should able to follow them rather than develop and create them. When it came to the life cycle of data, faculty had different perspectives. One believed that students did not necessarily have to understand the life cycle to manage the data. Another cited the data life cycle as the reason students lacked skills: they did not see the full picture of why data management and organization becomes important further in the data life cycle. Another faculty member maintained that it was important for students to understand the entire process so that they can backtrack if a mistake is made.

Student Responses

Students rated *data management and organization* skills as the highest competency in terms of importance. In general, the students described the processes of data management and not necessarily the reasons behind it. For example, most students kept copies of their data in

TABLE 3.5 *Faculty and Student DIL Competency Ratings of Importance: Data Preservation*

Competency-related skills:	Recognizes the benefits and costs of data preservation Understands the technology, resources, and organizational components of preserving data Utilizes best practices in preparing data for its eventual preservation during its active life cycle Articulates the potential long-term value of own data for self or others and is able to determine an appropriate preservation time frame Understands the need to develop preservation policies and is able to identify the core elements of such policies
Additional skills:	None
Faculty and student ratings:*	Faculty average = 3.57 Student average = 3.75

*Ratings based on a 5-point Likert scale: 5 = essential; 4 = very important; 3 = important; 2 = somewhat important; 1 = not important.

multiple locations, but the ad hoc methods of saving created confusion rather than security. Almost all students stated that they learned data management skills through trial and error. They learned through word-of-mouth about standards for managing and organizing their data, if they existed at all. Of the 15 students, 9 mentioned that there were no formal policies or that they did not know of any in place for managing the data in their lab (2 students did not respond to the question). Even those students working in labs with policies were unaware of formal standards in the discipline. The students recognized organization of data as an issue recognized for day-to-day tasks. For example, it was difficult for one student to locate particular files. That student reported occasionally needing to go back and rerun coding to find the authoritative version.

Data Preservation

Table 3.5 summarizes the results of our interviewee responses regarding the *data preservation* competency.

Faculty Responses

Depending on context, data preservation was considered either "essential" or not a major concern for faculty. Faculty whose work included sustainability of results over time tended to view preservation of their data as a priority. Other faculty saw the importance of preservation in theory, but did not necessarily see the need to take action to preserve their data. Faculty noted a lack of student knowledge or interest in this area. One faculty member mentioned a need for more resources to tell students about current best practices. Some faculty reported that they themselves did not have strong knowledge in this area. One rated this competency as both "important" and "I don't know," as he felt he did not fully understand data preservation. Another faculty member reported that since technology changed so quickly, some of the data would become obsolete quickly.

Student Responses

Many of the students were unsure of a long-term use for their data. Students gave a range

of responses when asked how long their data set should be preserved (see Table 3.6).

The length of preservation of data differed among the labs. For example, the students in the natural resources lab recognized the unique quality of their research and their role in supporting long-term research, and answered "indefinitely" to the question. Students in the agricultural and biological engineering lab were generally less certain of the long-term value of the data. Four of the five students responded either "less than 3 years" or "I don't know" to the question. There was some uncertainty about what was being done to preserve the data in the civil engineering lab. Two students indicated that no steps were being taken to preserve the data, one indicated that steps were being taken, and one did not know. Overall, students believed that the principal investigator, others in the lab, or a data repository handled data preservation.

TABLE 3.6 *How Long Should Your Data Set Be Preserved? (n = 17)*

Student Response	Number of Respondents
I don't know	4
Less than 3 years	2
10–20 years	2
20–50 years	3
50–100 years	1
For the life of the bridge being studied	1
Indefinitely	4

Data Processing and Analysis

Table 3.7 summarizes the results of our interviewee responses regarding the *data processing and analysis* competency.

Faculty Responses

Data processing and analysis is considered a direct component of conducting science in most disciplines; therefore it received the highest rating of importance by faculty. Overall, respondents viewed this competency as critical for students to avoid mistakes in evaluating data and to gain efficiency in their work. Several faculty mentioned that students were unfamiliar with processing and analysis tools in the lab as well as within their discipline.

Faculty estimated that their students' skill levels in this competency ranged from "not systematic" and "inefficient" to "highly experienced" upon entering the program. One faculty member described students as good in this area, but not necessarily efficient, meaning that it took students longer than it should to perform tasks. Potential resources for graduate students included workshops and classes, but peer-to-peer learning was noted as most influential. Another faculty member responded that he did not typically teach these skills because students absorbed the material better by engaging with it themselves—even though they may fail repeatedly.

As with many of the competencies, the nature of training depends on local and disciplinary practices and culture. There was an emphasis on developing processing and analysis skills and critical thinking through personal engagement with the data and tools. Some of the pathways to skill acquisition mentioned were peer-to-peer and advisor contacts; formal courses, such as statistics; and self-teaching/trial and error.

Student Responses

As with faculty, students recognized that these skills were generally at the core of scientific practice in their domains. One student from

TABLE 3.7 *Faculty and Student DIL Competency Ratings of Importance: Data Processing and Analysis*

Competency-related skills:	Familiar with the basic data processing and analysis tools and techniques of the discipline or research area Understands the effect that these tools may have on the data Uses appropriate workflow management tools to automate repetitive analysis of data
Additional skills:	None
Faculty and student ratings:*	Faculty average = 4.63 Student average = 4.35

*Ratings based on a 5-point Likert scale: 5 = essential; 4 = very important; 3 = important; 2 = somewhat important; 1 = not important.

the ecology lab commented: "One of the—I think—biggest mistakes that people make in our field is improperly analyzing data." Students indicated that they were asked to perform a wide variety of tasks in processing and analyzing data. Several students reported teaching themselves to use tools to perform these tasks. Statistical programs dominated the list of tools that students described (R, SPSS, SAS), as did Microsoft Excel. In addition, they described a variety of other programs and tools for collecting and transforming data specific to the particular research domain and project, including ArcGIS, data loggers, ENVI for analyzing Landsat images, MATLAB, and various coding languages such as Python and C++.

Data Quality and Documentation

Table 3.8 summarizes the results of our interviewee responses regarding the *data quality and documentation* competency.

Faculty Responses

Many faculty felt that their students knew to check for any discrepancies in their data to resolve issues before analysis; however, faculty did not express much confidence in their students' abilities to do the job well, nor to document the steps taken. One interviewee commented that it was "very hard to motivate students to write documentation," mostly because the students' focus was not on reproducibility, but on getting the work done and graduating. Faculty described self-documentation of code (a log of commands used and the parameters) as being important so that students could reproduce results. Another faculty member cited that a lack of tools for automating the process was a real challenge. This interviewee also noted that students consistently found themselves more concerned with the outputs of an experiment rather than the steps taken to get to the outputs. Still another faculty interviewee was confident that students were learning the skills needed to write the methods section of a paper, but that there was not enough documentation concerning the research process itself. This interviewee felt that students were overconfident when it came to artifacts and corruptions, and that they generally thought that their data was in good shape. One of the labs used error-checking procedures to ensure that measurements fell within known boundaries. The students in this lab participated in basic data quality checks, which included steps to ensure

TABLE 3.8 *Faculty and Student DIL Competency Ratings of Importance: Data Quality and Documentation*

Competency-related skills:	Recognizes, documents, and resolves any apparent artifacts, incompletion, or corruption of data Utilizes metadata to facilitate an understanding of potential problems with data sets Documents data sufficiently to enable reproduction of research results and data by others Tracks data provenance and clearly delineates and denotes versions of a data set
Additional skills:	None
Faculty and student ratings:*	Faculty average = 4.63 Student average = 4.12

*Ratings based on a 5-point Likert scale: 5 = essential; 4 = very important; 3 = important; 2 = somewhat important; 1 = not important.

that measurements were not out-of-bounds. Five out of the seven faculty we interviewed reported using some kind of version control practices in the lab, whether a specific system such as Subversion (SVN) or SharePoint, or file naming practices that included the version.

Student Responses

Overall, the students were aware of and/or participated in quality control steps. Out of 16 students, 14 felt that they created a sufficient amount of documentation for someone with similar expertise to understand and use their data (1 student did not provide a response). However, this may reflect one faculty member's assertion that students were overconfident in this area. Students in the computer engineering program were aware that this is an area that could benefit from "drastic improvement" (in the words of 1 student), but they also reported that their faculty advisor stressed documentation of the steps taken during research. For them, logging of calculations, thoughts, and the *entire* research process began early. These students were also more likely to use versioning software; students in ecology and natural resources were more likely to use file naming strategies for versioning. They learned these skills through trial and error, from peers, and from the principal investigator. All 16 of the students who provided a response planned to leave a copy of their data with their advisor after they graduate.

Data Visualization and Representation

Table 3.9 summarizes the results of our interviewee responses regarding the *data visualization and representation* competency.

Faculty Responses

Faculty saw data visualization and representation as a critical competency for students to master. They identified a need for more advanced instruction for students to learn how to create effective, and ethical, graphical representations of data. Several of the faculty reported that students learned the mechanical aspects of using visualization tools, but were not as skilled in knowing what makes a good visualization. As one faculty member stated, "visualization is communication." Students also struggled in making use of representations to evaluate the quality of their data or to "impact a specific decision."

TABLE 3.9 *Faculty and Student DIL Competency Ratings of Importance: Data Visualization and Representation*

Competency-related skills:	Proficiently uses basic visualization tools of discipline Avoids misleading or ambiguous representations when presenting data in tables, charts, and diagrams Chooses the appropriate type of visualization, such as maps, graphs, animations, or videos, on the basis of an understanding of the reason/purpose for visualizing or displaying data
Additional skills:	Understands the mechanics of specific data visualization software programs
Faculty and student ratings:*	Faculty average = 4.63 Student average = 4.35

*Ratings based on a 5-point Likert scale: 5 = essential; 4 = very important; 3 = important; 2 = somewhat important; 1 = not important.

Faculty reported that students received little to no formal training in this area as graduate students. Instead, students used the skills they acquired from undergraduate course work with their intuition to create visualizations and representations of their data. There were some exceptions. One faculty member recommended a book on the topic to incoming students. Another faculty member taught advanced techniques in the lab.

Student Responses

Student responses indicated a general recognition of the importance of data visualization to convey their findings in publications and other venues. All 17 of the students we interviewed indicated that they generated visual representations of their data. Several students mentioned the need to connect their work to their intended audiences. One student mentioned that "it's pretty much impossible to interpret the data without turning it into something." Students reported informal training on data visualization—advisors, lab mates/peers, and online help were resources for learning. Students mentioned a desire for software-specific instruction for creating their data visualizations in R, MATLAB, Python, GMT, ArcGIS, Excel, SPSS, GIMP, and SigmaPlot.

Databases and Data Formats

Table 3.10 summarizes the results of our interviewee responses regarding the *databases and data formats* competency.

Faculty Responses

Faculty stated that students needed competency with databases and data formats but that their abilities were generally underdeveloped. Faculty gravitated to the "databases" elements of this competency rather than the more general "data formats" aspects. This may be due to the order in which we presented our information; however, it can also be inferred that not every faculty member interviewed employed databases in his or her work. Not surprisingly, those who did tended to give a higher overall rating of importance to this competency than those who did not.

Of the faculty who discussed databases, most mentioned understanding how to query databases as an important skill for students. Any faculty thoughts and concerns about

TABLE 3.10 *Faculty and Student DIL Competency Ratings of Importance: Databases and Data Formats*

Competency-related skills:	Understands the concept of relational databases and how to query those databases Becomes familiar with standard data formats and types for discipline Understands which formats and data types are appropriate for different research questions
Additional skills:	Understands how to maximize performance of databases based on own design
Faculty and student ratings:*	Faculty average = 3.71 Student average = 3.88

*Ratings based on a 5-point Likert scale: 5 = essential; 4 = very important; 3 = important; 2 = somewhat important; 1 = not important.

databases were generally shaped by the way that they themselves made use of them in their labs. For example, the natural resources faculty member commented that without the use of databases, it's as if his data does not exist. In contrast, the agricultural and biological engineering faculty member was striving to incorporate all of the lab's data sets into a database and noted that both he and his students needed to spend more time learning about the capabilities of databases. Some fields offer courses in databases, and faculty expect that students take these courses and to know how to work with databases prior to joining the lab. The faculty we interviewed from fields in which such courses are not offered speculated that students acquired skills by working with others, rather than through formal classroom experience.

Student Responses

Students handled a variety of data formats in their respective labs. The vast majority of students used Microsoft Excel or .csv files, as well as ASCII text file formats. Other data formats mentioned were Microsoft Access databases, MATLAB files, images (TIFF and JPEG), raster data, SPSS files, SigmaPlot, and NetCDF, as well as the programming languages C and C++. Students tended not to focus on the data formats in the interviews. Therefore, they did not discuss larger issues in formatting data and databases in depth.

Discovery and Acquisition of Data

Table 3.11 summarizes the results of our interviewee responses regarding the *discovery and acquisition of data* competency.

Faculty Responses

Overall, faculty rated discovery and acquisition of data lowest of the 12 competencies. The assignment of importance to these skills seemed to align to the degree to which the individual and team used external data for research. Two of the faculty we interviewed indicated that the data they used were generated entirely in their labs, and they assigned a lower rating to this competency. Others indicated that external data might be brought into the lab to compare with or augment the data they generated. Or they might support an analysis done in the lab. Faculty used external data from sources such as the Census

TABLE 3.11 *Faculty and Student DIL Competency Ratings of Importance: Discovery and Acquisition of Data*

Competency-related skills:	Locates and utilizes disciplinary data repositories Evaluates the quality of the data available from external sources Not only identifies appropriate external data sources, but also imports data and converts it when necessary so it can be used locally
Additional skills:	Understands and navigates data use agreements for reuse of data sets from external sources
Faculty and student ratings:*	Faculty average = 3.57 Student average = 4.12

*Ratings based on a 5-point Likert scale: 5 = essential; 4 = very important; 3 = important; 2 = somewhat important; 1 = not important.

Bureau, SourceForge, and repositories of geospatial data.

Faculty thought that student skills were highly variable in this competency. They believed that students acquired skills through trial and error and consultations with advisors and peers. No dominant theme emerged across faculty responses, but some valued the ability to evaluate data quality and have an "appropriate level of skepticism of outside data sources." Some faculty thought that locating and using data sources, if necessary, was an easily acquired skill.

Student Responses

This competency was highly rated overall by students despite a lack of experience for some with locating and using data from external sources. Students reported that their skills were developed primarily from consultations with peers and advisors. Students' experiences in acquiring data varied. Some found data that had been well documented, thus making it easy to understand and use. Others noted that it was difficult to understand the external data they had acquired or the data used different measurement scales that had to be converted. Overall, 14 out of 17 students made use of data acquired outside of their lab. The major data repositories used by students were more varied than those listed by faculty. In addition to geospatial data repositories and SourceForge, students used the Environmental Protection Agency, the National Oceanic and Atmospheric Administration, Oregon State University's PRISM Climate Group, and the U.S. Department of Agriculture's Soil Survey Geographic (SSURGO) databases.

Seven out of the 17 students inherited data generated from others, reporting both positive and negative experiences in the transition. A student in computer engineering mentioned doing literature reviews as a means of searching for code.

Ethics and Attribution

Table 3.12 summarizes the results of our interviewee responses regarding the *ethics and attribution* competency.

Faculty Responses

Few faculty commented on the "misrepresentations of data" component of this competency, focusing instead on the citation, intellectual

TABLE 3.12 *Faculty and Student DIL Competency Ratings of Importance: Ethics and Attribution*

Competency-related skills:	Develops an understanding of intellectual property, privacy and confidentiality issues, and the ethos of the discipline when it comes to sharing and administering data Acknowledges data from external sources appropriately Avoids misleading or ambiguous representations when presenting data
Additional skills:	Identifies what data not to show for privacy purposes
Faculty and student ratings:*	Faculty average = 4.38 Student average = 4.35

*Ratings based on a 5-point Likert scale: 5 = essential; 4 = very important; 3 = important; 2 = somewhat important; 1 = not important.

property (IP), privacy, and confidentiality elements. Citing data was rated as "essential" to "very important" but faculty stated that their disciplines lacked standards for citing data. Most felt that students were good enough at citing data. One of the faculty members felt that ethics and attribution were discussed consistently in the lab and at the university and believed that students recognized that ethics extended beyond literature and included data sets. Two of the faculty felt that students cited outside sources sufficiently. One of them noted that students may not know how to cite a data set versus a piece of literature, and he himself didn't know of a disciplinary standard for citing data.

Several faculty noted that graduate students received ethics training either at the university or departmental level. The majority of the faculty noted that the question of who owned the data is "somewhat shaky" or "up in the air." One of the faculty members we interviewed felt that ethics training adequately covers privacy and IP issues, but more detailed, practical instruction for handling sensitive data is necessary. Another stated that students needed to understand the differences between copyrights, trademarks, and patents.

Student Responses

Several students reported citing the research paper associated with a data set rather than a data set itself, although many of the graduate students interviewed (11 out of 17 students) expressed a general feeling of being competent at citing data. It is encouraging that 11 students reported receiving training or instruction for ethics and IP issues, although they had mixed opinions about the usefulness of the training about data. Of the 17 students interviewed, only 3 indicated that they had a good understanding of their university's policies on research data, which echoed the faculty's statements on the need for more substantive graduate education in this area. One of the computer science students mentioned that the lab sought software code with open GNU or PSD licenses to ensure that they could properly use code generated by others. This aligned well with the faculty assertion that it was very important that these students understood issues with IP and copyright, trademarks, and patents. Of potential concern was that one student asserted that she didn't need to cite external code that she consulted but never used outright. About half of the students interviewed were not aware of any journals that

TABLE 3.13 *Faculty and Student DIL Competency Ratings of Importance: Metadata and Data Description*

Competency-related skills:	Understands the rationale for metadata and proficiently annotates and describes data so it can be understood and used by self and others Develops the ability to read and interpret metadata from external disciplinary sources Understands the structure and purpose of ontologies in facilitating better sharing of data
Additional skills:	Individuals who publish research must be ready at any point to answer questions from others regarding the data set
Faculty and student ratings:*	Faculty average = 4.57 Student average = 3.88

*Ratings based on a 5-point Likert scale: 5 = essential; 4 = very important; 3 = important; 2 = somewhat important; 1 = not important.

might accept data sets for publication or as supplements to a journal article.

Metadata and Data Description

Table 3.13 summarizes the results of our interviewee responses regarding the *metadata and data description* competency.

Faculty Responses

Faculty described students as barely proficient or worse in the area of metadata and data description, and most felt that this was an area that needed improvement. Nearly every faculty member interviewed (seven out of eight) reported that the amount of documentation and description that their graduate students currently provided was not sufficient for someone outside of the lab to understand and make use of the data. Three of the faculty reported that they themselves had some trouble understanding and making use of the data because of the lack of description. One of the faculty felt that this competency was of primary importance and that much could be gained by addressing the need; he expressed personal interest in learning more because he was unsure of the meaning of the term *metadata* and felt that a lack of knowledge in this area could be damaging. Another stated that "currently, researchers spend more time doing the work than explaining the work [they] are doing." For ongoing projects in one of the labs in which students pass code to other students each semester, the faculty member stated that current documentation was "definitely" not enough for someone outside of the lab to understand and make use of the data. Faculty considered this to be a major issue during project transition between semesters.

As an artifact of the research process, data sets are reflections of the decisions and actions made consciously or unconsciously by humans.

Student Responses

Out of the 17 students interviewed, 12 were familiar with the concept of metadata, though most stated that they had not received any formalized training. Some actually provided an inaccurate definition when pressed to explain it. (Two confused it with meta-analysis.) Student knowledge of metadata evolved from past projects, trial and error, and even past work in industry at least for one graduate

student. For example, a natural resources graduate student explained that her method for describing data had been learned through a "personal coping strategy," meaning, through trial and error. One graduate student familiar with metadata noted that the metadata he creates often is not detailed because he "doesn't have enough time." Several students reported no trouble understanding the metadata that accompanied the external data they have used. None of the students reported using a metadata standard, although one student applied a standardized taxonomy.

CONCLUSION

Overall the DIL competencies were an effective means of exploring the environments and needs of our faculty partners and their students. The DIL competencies were not intended to serve as a universally applied set of skills or as prescriptive standards. The DIL competencies will continue to evolve as we learn more about disciplinary and local practices. Chapter 10 addresses future directions for developing the DIL competencies.

We observed many commonalities between faculty and students from different fields of study and from different academic institutions. Conducting interviews informed not only our respective DIL programs but also our collective understanding of the environments in which research data are generated, administered, and utilized. As an artifact of the research process, data sets are reflections of the decisions and actions made consciously or unconsciously by humans. Understanding the environments, challenges, and needs of the people who work with data is an integral part of developing educational programs about data. The next section of this book presents the work of the five DIL project teams, describes the specific findings from their interviews, and their responses to the findings. These case studies illustrate how important the interviews were to the success of the DIL project.

NOTE

Portions of this chapter are reprinted from Carlson, J., Johnston, L., Westra, B., & Nichols, M. (2013). Developing an approach for data management education: A report from the Data Information Literacy project. *International Journal of Digital Curation, 8*(1), 204–217. http://dx.doi.org/10.2218/ijdc.v8i1.254

REFERENCES

Gullikson, S. (2006). Faculty perceptions of ACRL's Information Literacy Competency Standards for Higher Education. *Journal of Academic Librarianship, 32* (6), 583–592. http://dx.doi.org/10.1016/j.acalib.2006.06.001

McGuinness, C. (2006). What faculty think: Exploring the barriers to information literacy development in undergraduate education. *Journal of Academic Librarianship, 32*(6), 573–582. http://dx.doi.org/10.1016/j.acalib.2006.06.002

Weiner, S. A. (2014). Who teaches information literacy competencies? Report of a study of faculty. *College Teaching, 62*(1), 5–12. http://dx.doi.org/10.1080/87567555.2013.803949