圖資與檔案學刊
8:1=88 (June ’16) 1-27
DOI: 10.6575/JILA.2016.88.01

# Enhancing Scientific Data Literacy in College Students: Experience and Lessons Learned

# 提高大學生的科學數據素養：經驗與反思

Jian Qin *
秦　健

John D’Ignazio **

## 【摘要 Abstract】

Data literacy education is a new service territory for many academic libraries. This paper reports our experience from a two-year Science Data Literacy (SDL) project through four areas of activities: (1) survey on faculty’s perceptions and practices in data management, (2) design of SDL learning modules, (3) delivery of the course, and (4) assessment of learning outcomes. We found from the faculty survey that there was a low level of awareness of research data management importance, methods, and tools in general and that the data management practice is associated with the size and complexity of the data produced by their research. Science data literacy training was difficult to be integrated into formal curricula due to their structure, even though the need for such training was on the rise. The lessons learned from the two-year project and how libraries and library and information science education might turn this challenge into opportunities are discussed.

數據素養教育在很多大學圖書館是一個新的服務領域。本文報告我們所做的科學數據素養項目中的四個主要方面的活動：（1）對教師的數據管理意識與事件方面的問卷調查，（2）

* Professor, School of Information Studies, Syracuse University
E-mail: jqin@syr.edu

** Assistant Professor, Department of Information Science, College of Computing & Informatics, Drexel University
E-mail: dignazio@drexel.edu

數據素養教育的課程模塊，（3）課程的教學，（4）學習成效的評估。從對教師問卷調查中我們發現教師對研究數據管理的重要性、方法以及工具普遍認知不足，這種認知的程度與他們所產生的研究數據的數量大小與複雜程度有關。由於科學專業的課程設置結構的原因，即使需求在增長，科學數據素養的訓練融入本科正規課程中有一定困難。文中討論了從該項目的執行過程中獲得的經驗、反思以及對圖書館學資訊科學教育的挑戰。

## 關鍵詞 Keyword

科學數據素養 數據素養教育 科研數據管理

Science data literacy；Data literacy education；Research data management

# Background

The practice of science has changed in the last three decades due to the rapid development of information and communication technologies and massive increases in computing capacity, made manifest by the Internet. As the International Council for Science (ICSU) describes in its five-year strategic plan, there are more scientific data and information openly available than ever before. This environment enables scientists around the world access to the most up-to-date data and information from his or her desktop. “Secondary analyses of data, and the combining of data from multiple sources, are opening up exciting new scientific horizons. Scientific publication practices are changing rapidly.” (International Council for Science [ICSU], 2005, p. 17)

The changes in scientific research practice not only created challenges for preparing science students with data literacy, but also generated a great demand for a workforce that is trained in both scientific disciplines and data management. It is not uncommon in job notices these days that candidates for data management jobs are required to have knowledge and skills in data science methodologies and applications. Such requirements delineate a typical situation in today’s scientific research in which scientific data literacy goes beyond the skills in using the data: an awareness and experience of data formats, organization schemes, tools, communication, conversion, and manipulation is becoming a valuable, indispensable part of the qualifications for future workforce.

In the last 8 years, we conducted two projects related to Scientific Data Literacy (SDL) training and e-science librarianship, the former of which was funded by U.S. National Science Foundation and the latter by the Institute for Museum and Library Services (IMLS). The goal of the SDL project was to create an SDL course for undergraduate and graduate students to learn the fundamental concepts in science data and support the use of data in the course of scientific inquiry; as one of the outcomes, the SDL course would help prepare students majoring in science and technology for a career in science data management. The e-science librarianship project was to recruit students with a science background to librarianship and to research and develop a new curriculum that responds to needs for management of new and different types of digital resources, at amounts previously unimagined, for long-term access and use. On the education side, both projects involved recruiting students, delivering the new courses and evaluating learning outcomes. Research was also part of the projects in both cases.

Data literacy training is still a field to be explored, especially in the data-driven science and decision making environment. Many academic libraries in the U.S. initiated data literacy programs, but so far difficulties exist in making data literacy part of their curriculum. This paper will focus on reporting the experience and lessons learned from the scientific data literacy project, with the details from the course development and outcome assessment. Implications of the findings from this project are also discussed at the end of this paper.

## Understanding Faculty Data Practices

### *Survey Design*

This survey was designed as a census of the relevant campus faculty to understand their data practices. Although several surveys on researchers' data practices have been reported in literature (Akers & Doty, 2013; Whitmire, Boock, & Sutton, 2015), there was little in published research on this topic at the time of our survey design back in 2008. The project team took a pragmatic approach in crossing disciplinary boundaries in order to gather the variety of data management practices in departments within Science, Technology, Engineering, and Mathematics (STEM) disciplines across the home institution. Departments were identified that reasonably fell within the rough boundaries provided by the STEM category amalgam; SDL staff erred on the side of including the most number of researchers likely to be accumulating and working with primary datasets.

Based on the goal of this data collection, we divided the questions into three categories: attitudes about and use of data, management of data, and demographics. The survey was created and pilot-tested on members of the SDL project advisory board who matched the target population. Terminology had to be negotiated and definitions were provided to orient concepts from information science to the scholars' research process. For example, we eliminated the word "metadata" from the questionnaire and added an inclusive definition of data from a National Science Board report (2005) in the introduction to the survey: "any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc." Demographic questions particular to the discipline-focused and hierarchical environment of the academic community were taken from the Higher Education Research Institute's faculty performance survey (2004). As per Janes' (1999) helpful advice on survey construction, this demographic data-gathering section was placed in the end.

The iterative process of question phrasing and grouping led to a refined instrument designed to capture data management practices. The four-part, web-based survey featured Likert-scale agreement questions related to attitudes, practices, and experience with research data. Two provocative questions about data management in the respondent's discipline were designed not only to motivate participants to express their views, but also to help establish the presupposition of the questionnaire that the respondent was a data producer (Martin, 2006). This rather brute force technique was designed to concentrate the mind of the respondents on their data management practices, but it may have had unknown effects on the target population. An initial branching question may have been more effective at weeding out members of the STEM departments who, due to a teaching, practical, or theoretical orientation, did not manage data. In the meat of the survey, branching questions would also have allowed researchers to record more than one management or preservation practice, but the instrument provided for this purpose enabling for the participant to mark multiple options per response was appropriate to simplify the survey design. Textboxes and open-ended questions in each section were meant to elicit a range of practices with data as well as overall reactions provoked by the 25 questions.

Syracuse University, in combination with the symbiotically connected campus of the SUNY College of Environmental Science and Forestry (SUNY-ESF), has over a thousand full-time faculty members; the SDL survey was administered to 362 faculty members from STEM departments at the two institutions. Respondents received a movie ticket coupon as consideration for their time. Possible participants were identified via information posted on school and department websites and then contacted via email solicitation, notification and up to two reminders. Faculty members were given the opportunity to opt out—many did citing current retirement status or lack of involvement in research. As it was a local census using a MySQL-based token response system, anonymity was not an option, so entries are being kept in confidence. The questionnaire is provided in Appendix A.

## *Faculty Perceptions of and Practices in Science Data Management*

Of those who participated resulting in a 30.7 percent response rate (111 responses out of 362 faculty members contacted), problems that might affect the census results include a lack of alignment of the respondent with data-producing aspects of STEM research pursuits, such as a solely theoretical orientation in physics and mathematics, or a social orientation in the case of geography and information science. Additionally, faculty who handle different types of data per

research project, faculty organized in research groups, and faculty use of research assistants as day-to-day handlers of data are conditions that may have interfered with an accurate and complete perception of data management practices from the survey responses obtained. The 111 responses came from participants with a wide range of disciplines that resulted in smaller numbers of responses in most of the disciplinary fields, which was not feasible for inferential statistical analysis. Therefore, the analysis of survey results will be primarily descriptive due to the limitation of data.

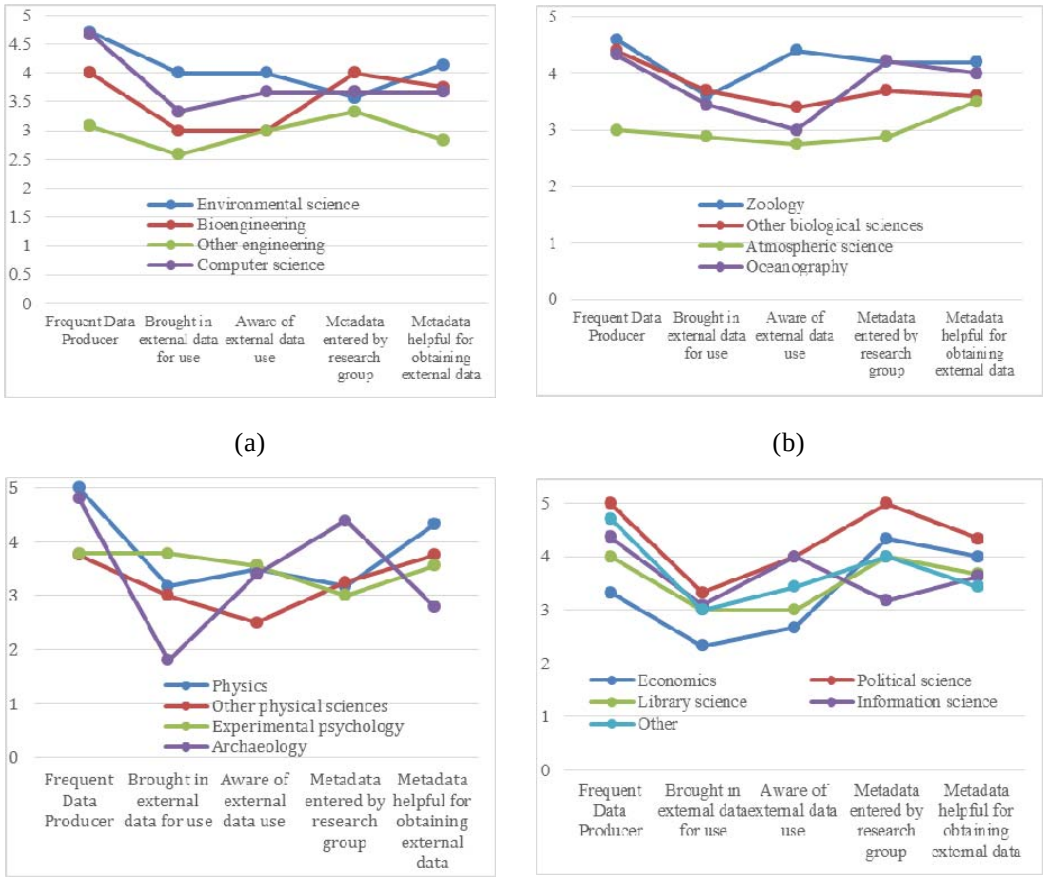


Figure 1　Relationship between discipline, data production & use, and metadata creation & use
(3 or more responses identified by discipline, 5-point agreement scale, high value indicates more agreement)

Figure 1 indicates some of the variety encountered in data management practices throughout the STEM departments. A value closer to five indicates a strong agreement that the researcher respondent was a frequent data producer, used data prepared by another researcher,

was aware that the digital data may be used by another researcher outside his or her own group, prepared metadata of some kind for the internally produced datasets, and found metadata entries helpful when obtaining data for use from external research groups. All entries could have been the respondent answering as a representative of his or her research group. While respondents in most disciplinary fields rated high as frequent data producers, responses to the four categories other than frequent data producer varied between disciplinary fields. For example, archaeology was a field in which researchers were most unlikely to bring in external data for use (Figure 1-c). Social sciences (Figure 1-d) showed a quite consistent pattern among all five categories, but Figure 1-a illustrates differences in bringing external data for use at different magnitudes with other engineering being the lowest, environment science the highest, and bioengineering and computer science in between.

Figure 2 shows the relation in responses between those researchers who worked with a certain size dataset and their perception of the effect of data management practices on their discipline's progress. Researchers who operated with larger datasets appeared more confident about their discipline's data management practices.

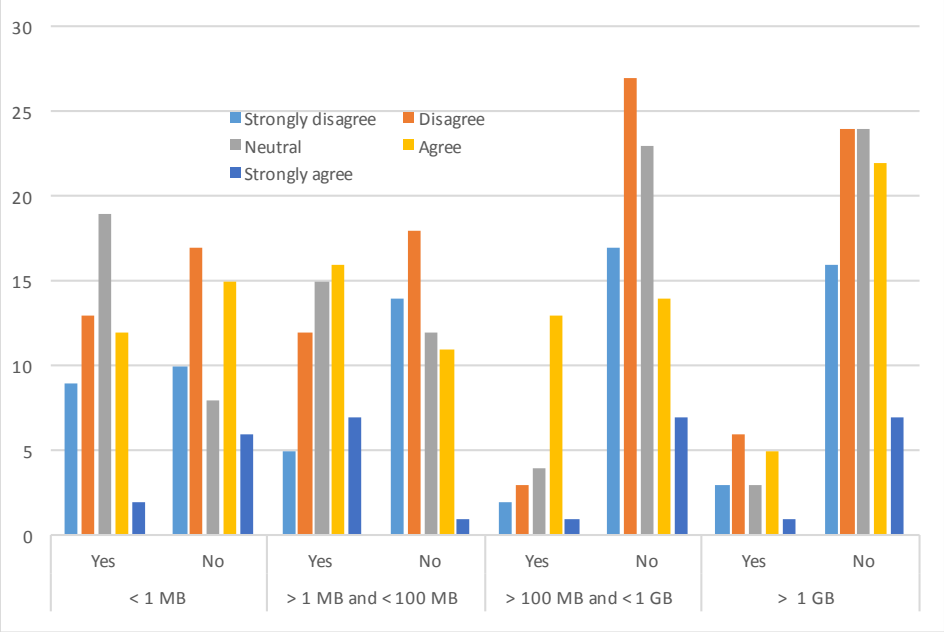


Figure 2 Relationship between current data management practice and the size of data files
(Selection of data file size X, 5-point agreement scale regarding practices limiting disciplinary advancement)

Data management actions (or inactions) can have an impact on disciplinary information loss. Participants who routinely performed cleaning, conversion, merging, calculation, and visualization with their data agreed that inadequate data preservation practices had negative impact on their discipline (Figure 3). It appears that researchers involved in advanced data activity such as calculation and visualization may be slightly more sanguine about their discipline's preservation practices.

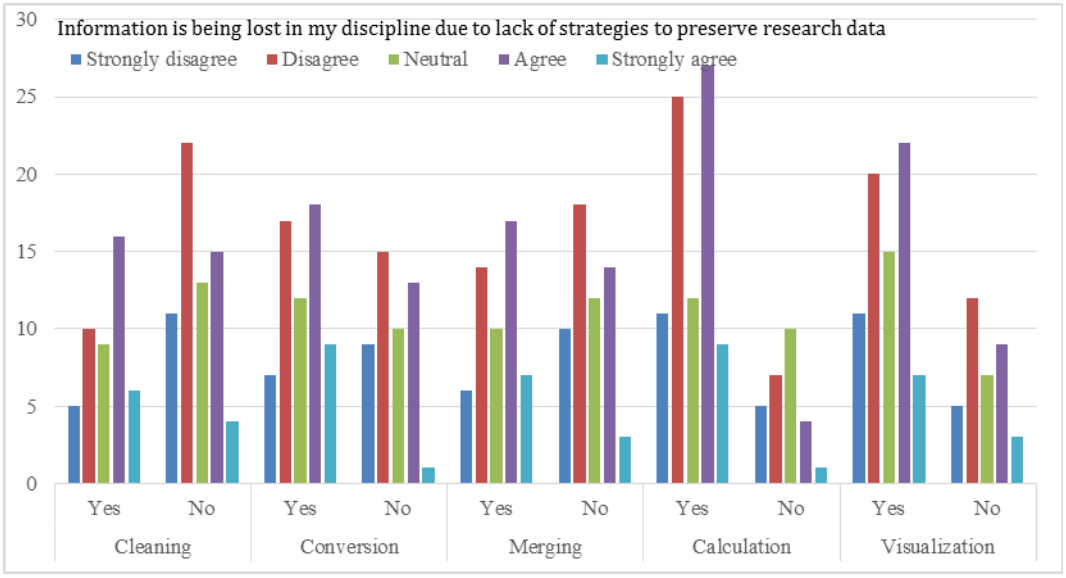


Figure 3　Impact of lack of data management strategies on data use and management tasks
(Selection of data management action X, 5-point agreement scale regarding disciplinary information loss)

Continuing analysis of the survey data showed the variations in management and preservation practice, according to faculty affiliation and position. This allowed mapping of local institutional attitudes and behaviors regarding data to those encouraged by major research initiatives at the disciplinary and national level. The results from this survey helped the SDL project team and the library and information science (LIS) field more generally understand how to integrate SDL education with STEM departments actively working with science data, as well as develop educational materials at an appropriate level to assist meeting the need for personnel with the skills and interest in science data management and preservation to support effective community and disciplinary use of these digital resources. On a more basic level, it helped provide information on how science and technology researchers obtain and manage their data as part of knowledge production and science communication processes.

# Design of the Course

With input from the survey, we focused the design of the science data literacy course on three major areas: content, pedagogy, and assessment.

## *Content Design*

We started the design process by learning about what other institutions are offering in data related topics. We compiled a table (Table 1) representing this environmental scan to compare the contents in these relevant courses. 2 of 20 courses in Table 1 were identified as the most closely related to the intent of the SDL project in terms of science data management. We adopted SDL as our course title because of the strong definition and associations with a practice, recognizing that these two courses had good coverage of data technology and tools. There was, however, a lack of topics in these courses on the evaluation of data quality, data publishing and repositories, and ethics and intellectual property issues in science data—some of which were represented in the data curation course. Most courses in Table 1 were designed for the graduate level.

Table 1

*Categories of Relevant Courses (UG=Undergraduate)*

| Course Category | Number of Institutions | Level | Focus |
|---|---|---|---|
| Computing tools | 1 | UG for non-majors | Computer-based tools useful for analysis and understanding of scientific data |
| Data analysis | 1 | Graduate | Techniques of exploratory data analysis using scripting, text parsing, structured query language, regular expressions, graphing, and clustering methods to explore data |
| Data curation | 1 | Graduate | An overview of theoretical and practical problems in data curation and examination of issues related to appraisal and selection, long-lived data collections, research lifecycles, workflows, metadata, legal and intellectual property issues |
| Database systems | 3 | Graduate | Database design and implementation |
| Information representation | 1 | UG | Principles and techniques of organizing non-bibliographic information sources including unpublished and transitory materials such as archival and manuscript collections, business/office records, ephemera, and local databases |
| Information systems | 1 | Graduate | Cognitive work analysis framework and design and evaluation of information systems |

Table (Continued)

| Course Category | Number of Institutions | Level | Focus |
|---|---|---|---|
| Metadata | 9 | Graduate | Metadata standards, creation, retrieval, and management |
| Science data management | 2 | UG and graduate | Theory, techniques, and tools for managing heterogeneous scientific information, database architectures, and data models; information and technological literacy (IL & TL) necessary to succeed in a scientific vocation |
| Science information services | 1 | Graduate | Information resources and services in science and technology including primary and secondary publications, electronic text, image and numeric databases; user needs and communications patterns within the scientific community |

To achieve the goals of this project, for undergraduate students to understand the fundamental concepts in scientific data and to use the data for scientific inquiry, our analysis of the field indicated the contents needed to cover not only the technology used for managing science data but also analytical and evaluative skills for using data and/or providing data services. Using this rationale, we grouped data-related topics into three modules shown as below:

Table 2

*Topics Covered in the Science Data Management Course*

| Module1: Fundamentals of science data management | Module 2: Case studies and issues in science data management | Module 3: Use of science data |
|---|---|---|
| Research lifecycle, research data lifecycle, fundamentals about data, representation and organization of datasets, technologies and infrastructures for research data management, the concepts of data provenance and research reproducibility | Data use and management scenarios, developing a data management project, which includes domain data analysis, needs assessment, data policy development, long-term preservation of data, etc. | Data quality, data publishing and sharing, data citation, data analysis, data presentation, and ethics in using science data |

In the design of the course, we distinguished between managing data for use and managing data as a career. While the boundaries between the two are often blurry, this distinction allowed us to provide a balanced list of the topics for both data users and data managers. The full list of topics can be found in Appendix B.

## *Pedagogy*

The pedagogy of this course not just focused on individual skill and knowledge acquisition, but required shared learning and interdisciplinary teamwork opportunities due to the nature of the material. Students engaged in group work and interactions with both fellow classmates and faculty through guest speakers and interviews. Class and group discussions about data use were guided by the framework shown in Table 3, which is based on informational, explanatory, procedural, heuristic, and valuative aspects (Manduca & Mogk, 2002).

Table 3

*A Framework of Questions for Class Discussions to Understand Data Use*

| | |
|---|---|
| Informational | |
| Descriptive:<br>• What is it?<br>• What kind is it?<br>• Where is it?<br>• When is it?<br>• Who is it? | Operational:<br>• How does it work?<br>• What does it do? |
| Explanatory | |
| Causal:<br>• Why does it work that way?<br>• What is the reason for that? | Teleological:<br>• Why did he do that?<br>• What was the purpose for that? |
| Procedural | |
| • Methodological:<br>• What is done?<br>• What could be done? | Technical:<br>• How is that done?<br>• Is it done this way? |
| Heuristic | |
| • Investigative:<br>• What could we find out?<br>• How could we find out? | Speculative:<br>• What would happen if?<br>• What could happen if? |
| Valuative | |
| • Normative:<br>• Is it any good?<br>• How good is it? | Significance:<br>• What difference does it make?<br>• So what? |

*Note.* From *Using Data in Undergraduate Science Classrooms: Final Report on an Interdisciplinary Workshop at Carleton College* (p. 40), by C. A. Manduca, & D. W. Mogk, 2002, Northfield, MN: Carleton College Science Education Research Center.

In addition to the three topic-oriented modules, the course also had a two-part organizational strategy. The first part of the course featured more traditional classroom-based activity involving lectures with four exercises designed to instill data management skill proficiency. These exercises have a unique self-reinforcing quality in reinforcing the student's engagement with a domain of interest and developing a strong sense of what is involved in data management. Firstly, the students were asked to choose a resource on the web (or from a list of identified options) to examine and proceeded to a reverse engineering repository, describing its structure and the content of an example dataset it contains. They based the work on the next three exercises on this initial analysis, and could go back to the web resource to fill in gaps as they created a database design, compose query statements that fits a scenario where a scientist engaged with the repository to gain resources, and create a metadata or data record suitable for the repository contents.

The second part of the course took the students out of the classroom and into the science information environment in the form of an authentic project involving the actual data use of a professor or a research group in the field of science or technology. We arranged a list of scientists for the students, working in pairs, to contact, mostly made up of our advisory board. A majority of student groups took their own initiative and followed their interests to start projects with professors in the field of science or technology around campus, thus extending the reach of the SDL project in a dynamic, unplanned way. The sequence of activities for the student groups started with becoming familiar with their target scientist's work via their website and publications, which helped them prepare for an interview with the scientist. Conducting an interview was therefore included as a skill the course helped them develop.

The students converted the interview experience into a group presentation and accompanying report about their scientist's data and technology use. The next assignment was to apply the four exercises to some aspect of the scientist's data use or management environment. Finally, based on this design process, the student groups created a data or metadata management prototype solution appropriate to assist the scientist in their work. A concluding presentation and report helped the student teams synthesize their experience and share what they had learned and created. As with the exercise structure, each project step was built upon a previous activity and enriched by continued readings, discussions, and lectures.

## *Assessment*

A number of methods were used to assess student's performance and learn about their literacy and aptitude related to data. For academic performance assessment, we used (a) quizzes to test students' understanding of the basic concepts in science data management mostly obtained from lectures and readings, (b) exercises for students to practice necessary skills in carrying out data management projects that were modeled in lectures and case analyses, (c) presentations and group reports to allow students to share their thoughts and case analysis regarding data management and use scenarios, and (d) an authentic project to evaluate students' knowledge and skills in data management and use as well as creativity and originality in designing a data management prototype solution.

We also designed a pre- and post-course survey, available in Appendix D, to examine changes in students' perceptions and confidence on the subject before and after taking the course. The pre-course survey assessed the students in four areas: (a) experience with computer technology, including databases; (b) affinity for science-related topics; (c) comfort with teamwork; and (d) comfort with the course topic and learning objectives.

The pre-course survey therefore established a baseline of students' knowledge and technical skills as well as their perceptions related to science data. To measure the comfort with the course topic and learning objectives, the questionnaire design included the Perceived Competence Scales (PCS) (Williams & Deci, 1996; Williams et al., 1998). Open-ended questions were asked for specifics regarding the science domain they had exposure to or were interested in, experience with software tools related to working with data, an instance of problem-solving in a group, and expectations for the course.

Among the 14 (6 undergraduates and 8 graduate) students who took the course, most had moderately above average experience in technology and data management (mean=3.45 of a 5 scale, with a std. dev. 0.902). A majority of students came from the Master of Science in Information Management (IM) and in LIS programs, with two students from STEM departments. In response to question "I am familiar with a scientific discipline and its methods", students named astronomy, computer science, geospatial, mathematics, and biology. Most students also had experience in team work from other classes they took. When students were asked the question "What made you want to take this course?", the answers varied in a wide range:

1. (UG) “Never took a data course before.”
2. (UG) “Something different from other IST classes.”
3. (UG) “Seemed like the most interesting IST Management course.”
4. (UG) “Learn more about database structure.”
5. (UG) “I don’t know.”
6. (G) “Relevance to database.”
7. (G) “Possible career choice in data management and research.”
8. (G) “Learn some methods and technology.”
9. (G) “How to manage the data which I collect and generate.”
10. (G) “I have some interest in it & it is helpful for my future career.”
11. (G) “I was inspired by a guest lecture by Clifford Lynch in one of my classes on digital libraries and cyberinfrastructure.”
12. (G) “I like working with very large data sets collected by others or re-used from their original purpose.”
13. (G) “I want to be a science librarian.”

The post-course survey was designed for two main purposes: (a) assess the course outcomes by comparing students’ responses to questions to the baseline established in the initial survey, and (b) assess the instructional content and methods. To accomplish the latter purpose, the questions addressed the course components, such as readings and the project, in terms of whether the students found them useful and interesting. Open-ended questions were asked for specific areas of instructional effectiveness or ineffectiveness, students’ satisfaction or dissatisfaction, and science domain experienced. The post-course survey is discussed in the following section.

## Discussion of Findings

The course *Scientific Data Management* was offered three times during 2008-2011 before being regularized into the formal curriculum at the iSchool. The title of the course was changed later to *Applied Data Science* with a shift of emphasis to data analytics, and it became a popular course among students in both IM and LIS programs. The SDL project, the course design process and the learning outcome assessment offered some insights into the research data literacy training to college students.

### *Effect of Science Curriculum Structure on Enrollment*

Student enrollment in the scientific data management course proved to be a major struggle for the project. This issue came up as early as our kick-off meeting, when the advisory board members expressed concerns that the structure of undergraduate science curriculum could deter even the students who knew of and wanted to take the course. Given that the course's nature as a special topic was not listed to be part of any programs or specialties, the first step to surmount this problem was to raise awareness of the planned offering. We tried various ways to promote the course, including (a) an expert panel on science data management and use in the real world, (b) flyers posted to campus buildings and passed out in departments by advisory board members, (c) email distribution to STEM faculty members targeted as possible participants in our survey, (d) a campus mailing to survey participants and student advisors in STEM departments, and (e) visits to science classes for a short presentation and Q&A session.

After all these efforts, we managed to retain 14 students in the first offering, 10 students for spring 2008 and 8 students for spring 2009. While there are several factors contributing to the low enrollment, the current science curriculum structures seem to have a significant impact on the enrollment. Many of the science curricula are structured in a way that each undergraduate is mandated to complete a fixed number of required courses and electives. Working with faculty advisors to create a schedule where they can complete these required courses for graduation leaves little room for students to take cutting-edge and somewhat demanding courses such as the courses related to science data management, particularly for the juniors and seniors who are the best candidates to take this course.

Of the classes we visited in our outreach efforts due to sympathetic faculty members' responding to an email request and inviting us to describe the course at the beginning of their class period, and there seemed a strong interest in having their students take our course, including the contribution of their personal stories regarding the need for effective science data management. Results from our campus-wide census on this issue were consistent with this feedback gathered from the field. Sentiment on the need for attention to data in the sciences can be roughly divided by equal thirds: those who are neutral, those who think things are okay for now in their discipline, and those who think attention should be paid. Given the wide variety of STEM-related activity on campus, this does seem like a large amount of interest, and yet such an interest and relative consensus have yet to find an effective outlet for enhancing scientific data literacy education. Through our advisory board, we heard of one effort at the department

level to educate students about data management, but this seems an inefficient approach to an increasingly common problem.

## *Changes in Students' Perceptions and Aptitude*

At the beginning and the end of the first two offerings in this course, we conducted an evaluative survey targeting student attitudes and reaction to the course experience. Students enrolled in this course felt a closer affinity with science disciplines, but their relative comfort with technology and experience in databases and teamwork varied by the year (Table 4). This is reflective of the fact that more students enrolled in this course came from a science discipline in the first offering while a majority in the second offering specialized in a technology field. The drop in perceived competence in the second offering may reflect the relative inexperience of the instructor who was a Ph.D. student, but it may also reflect his insistence that they worked with data or metadata from a target science domain, not staying in comfort zones such as web design or writing papers.

Despite the change of teachers and the adjustments to the course, the reaction to the pace of the class stayed roughly the same. Students generally considered the assignments/project the most interesting and useful part of the course (mean score 4.44 out of 5), and the results corresponded to the positive comments about the value of the project experience. The high marks for the exercises, despite the complaints due to the ongoing adjustment and design of assignment specifications as the course went along in the second offering, indicated that the students felt they learned the necessary skills with the exercises.

Table 4

*Comparison of Student Attitudes about Science and Being a Data Manager*

| | First offering: Initial survey mean (std dev) | First offering: Final survey | Second offering: Initial survey | Second offering: Final survey |
|---|---|---|---|---|
| N | 14 (6UG+8G) | 9 (4UG+5G) | 11 (4UG+7G) | 8 (2UG+6G) |
| Science Affinity (up to 5) | 2.32 (1.08) | 3.30 (1.02) | 3.27 (1.13) | 3.75 (.46) |
| Comfort with Databases and Computers (up to 5) | 3.45 (.90) | 3.96 (.75) | 2.91 (.94) | 3.19 (.75) |
| Comfort with Teamwork (up to 5) | 3.29 (.89) | 4.11 (.93) | 2.82 (.90) | 3.0 (1.31) |
| Perceived Competence Scale (up to 7) | 5.61 (1.25) | 5.83 (1.15) | 5.68 (1.07) | 4.18 (1.04) |

We collected student comments as the responses to the open-ended questions asked in the post-course survey in both years. These can be compared with the Likert-scale responses of the rest of the surveys and also simply appreciated as honest reactions to the course, to science, and to their newly found appreciation for the role of science data management. More precise wording about the choice of activity between metadata or data level may lead to a better understanding. Example solutions are again called for, although these "canned" illustrative examples take much time to prepare and need to be applied across the science domains the students might choose, unless the science domains are restricted to allow more cogent teaching.

Table 5

*Students' Responses to the Course Content Questions*

| Three things I learned from this course: | Three things I wish I learned from this course: |
|---|---|
| (UG) XML Data Scheme use | (UG) I learned it all. |
| (UG) Deepened my database knowledge | (UG) Database creation |
| (UG) How IT and science can mix | (UG) I didn't really understand much of the scientific stuff, so explaining that more might help. |
| (UG) metadata is what can be used to describe data | (UG) more technical things |
| (UG) how to use query statements on asp.net platform | (UG) understanding the exact jobs of scientific data managers |
| (UG) importance of organizing information in a way that is useful and neat | (UG) more on how to code XML |
| (UG) Metavista | (UG) more ERD work |
| (UG) Data organization | (UG) more Access use |
| (UG) Database | (G) some software and techniques for metadata |
| (UG) Schema | (G) database design |
| (UG) Metadata standards | (G) more about repository design |
| (G) Metadata schemas | (G) database |
| (G) science data lifecycle | (G) metadata |
| (G) how to apply concepts to development | (G) more content on data curation processes |
| (G) creating and managing a database | (G) preservation of scientific datasets |
| (G) scientific research process | (G) how to create metadata with XML |
| (G) general metadata creating and management that are applicable in a scientific research project | (G) professional real-world scenarios (more) |
| (G) metadata | (G) how to create science data |
| (G) metadata standards (FGDC) | (G) I wish I learned more from other students. |
| (G) ER diagram | |
| (G) FGDC standard | |
| (G) E-R diagram | |
| (G) issues of data management | |
| (G) scientific data management is very complex | |
| (G) developing a data management system involves going through a series of steps and thought processes | |
| (G) different scientific disciplines alter metadata schemas to fulfill the discipline's needs | |

We also asked students to list three things they learned and they wished they had learned from the course (Table 5). The answers provided the insights into the instructional design and pedagogy in teaching the course.

Regarding the question “What is your favorite part of the course? Why?”, students’ comments are listed as below:

1. (UG) “The project: it allowed us to take what we learned and utilize it.”
2. (UG) “Making the database and website: It’s really fun to create something useful for other people, that makes their jobs easier.”
3. (UG) “Exercises 1-4: Helped out with database skills, ER-Diagram creations.”
4. (G) “Group project: got to apply concepts we learned.”
5. (G) “Listening to researchers describe their work gave a great overview of scientific issues that affect everyday life: It was the first time I was exposed to scientific research processes, which were enlightening at many points. It also introduced useful metadata management concepts that are applicable to my area of interest.”
6. (G) “Design a repository: it’s interesting.”
7. (G) “Fundamentals about science data & data management with case studies: easier to learn.”
8. (G) “Learning how to use new technologies: The lectures were helpful in learning how to use new programs like Oxygen XML editor. Once I learned the basics, I felt really comfortable using it and built up my confidence. I never used XML before this course.”

## *Data Literacy Education*

We learned several lessons from our experience in science data literacy education. The first one is the gap between what the e-science and e-research environment needs from the new science workforce and how much the STEM faculty have become aware of these expectations. This gap is reflected in a lack of response reflected in the science curricula to accommodating the changing e-science/e-research environment. Our outreach experience to science classes demonstrates the existence of such a gap. While offering the course on science data management and use helped mitigate the gap to some extent, we also realized that data literacy education would not be effective without the participation and support of STEM faculty. The one-course-for-all approach may need to be adjusted to be more flexible and accommodating to disciplinary idiosyncrasies. This leads to the second lesson learned on the appropriateness of

potential students/audience for the content covered.

STEM fields vary in terms of data scope, type, format, subject, and size. Even within the same field, the purposes and kinds of research can result in very different sets of data attributes. Realizing the differences among STEM fields, we tried to make each class interesting and easy to understand through using data examples from various science research fields. We observed, however, that among the students enrolled in this course, graduate students had much better comprehension of the material and intellectual engagement in the exercises than undergraduate students did. This prompted us to pair undergraduate and graduate students together in project groups to make sure each group had a good balance among the levels of intellectual maturity, the types of skills and the subjects' background.

These observations led us to believe that science data literacy training needs to be provided at different levels via different ways, and that the training needs to adapt to science disciplinary context, terminology, and workflow. At the undergraduate level, the goal would be to train future science workforce with a solid understanding and skill set in the issues of data management and use. Such training would be more productive if the science data literacy content can be incorporated into science curricula so that the data literacy topics and practices can be put in context, perhaps as a unit within already existing labs or courses.

While science data literacy skills are becoming increasingly important as the e-science/e-research environment evolves, administrative support at the university and college levels is critical for the success. This includes, among other things, the awareness about science data literacy among university and college administrators and accompanying support for experimental and interdisciplinary projects that incorporate data literacy training for undergraduate and graduate STEM students.

## Conclusion

The science data literacy project was a new attempt at the time to integrate data literacy training into undergraduate and graduate training. Around the time of this SDL project, the type of big data, data-driven research and decision making prompted the emergence of digital curation, data curation, and data science courses and programs, and also changed the landscape of data literacy training. Today data curation and data librarianship, mostly taking place in academic libraries, are common terms in professional publications and have become the new

frontier among the services provided by academic/research libraries. The lessons learned from this early project offered valuable insights into the demand for what kinds of data literacy are needed and how they can be more effectively delivered to college students and faculty members. Data literacy training is becoming an important part of the library services, which raises challenges for librarians to have the knowledge and skills to provide such training programs as well as library and information science education.

（收稿日期：2016 年 3 月 31 日）

## References

Akers, K.G., & Doty, J. (2013). Disciplinary differences in faculty research data management practices and perspectives. *International Journal of Digital Curation*, *8*(2), 5-26.

*Faculty survey* (2004). Los Angeles, Calif.: Higher Education Research Institute, University of California, Los Angeles. Retrieved August 07, 2007, from http://www.gseis.ucla.edu/heri/researchers/instruments/FACULTY/

International Council for Science. (2005). *Strengthening international science for the benefit of society: A strategic plan for the International Council of Science 2006-2011*. Retrieved from http://www.icsu.org/publications/reports-and-reviews/icsu-strategic-plan-2006-2011/icsu-strategic-plan-2006-2011.pdf

Janes, J. (1999). On Research: Survey Construction. *Library Hi Tech*, *17*(3), 321-325.

National Science Board (2005). *Long-lived digital data collections: Enabling research and education in the 21st century*. Washington, D. C.: National Science Board, National Science Foundation.

Manduca, C. A., & Mogk, D. W. (2002). *Using data in undergraduate science classrooms: Final report on an Interdisciplinary Workshop at Carleton College*. Northfield, MN: Carleton College Science Education Research Center**.**

Martin, E. (2006). *Survey questionnaire construction* (Research Report Series #2206-13). Washington, D.C.: U.S. Census Bureau.

Whitmire, A. L., Boock, M., & Sutton, S. C. (2015). Variability in academic research data management practices: Implications for data services development from a faculty survey. *Program: Electronic Library and Information Systems*, *49*(4), 382-407.

Williams, G. C., & Deci, E. L. (1996). Internalization of biopsychosocial values by medical students: A test of self-determination theory. *Journal of Personality and Social Psychology, 70*, 767-779.

Williams, G. C., Freedman, Z. R., & Deci, E. L. (1998). Supporting autonomy to motivate glucose control in patients with diabetes. *Diabetes Care, 21*, 1644-1651.

# Appendix A

## Questionnaire for Science Data Literacy: Faculty Survey

You are likely familiar with the rapid changes that computer and network infrastructure developments are having on your scientific research and the work practices of those in your discipline. The Science Data Literacy (SDL) project has been funded by the National Science Foundation to study how these developments are affecting you and the way you are managing the increasing amounts of data stored as digital files. We use the National Science Board's definition of data as "any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc." Principal Investigator and Associate Professor Jian Qin from Syracuse University's School of Information Studies (iSchool) invites your participation in a survey of the science, technology, engineering, and mathematics faculty at this institution.

The survey should take you no more than 20 minutes to complete, and to compensate you for your time, you'll receive a movie ticket redeemable at any Regal Entertainment Group theatre such as at Carousel Mall. Your entries will be kept in strict confidence and the survey data collected will be reported in aggregation only. We will make sure that any publication resulting from this survey will not provide a means for identifying you or your work. If you have any questions about this survey or the Science Data Literacy project, please contact SDL research assistant John D'Ignazio at jadignaz@syr.edu. We thank you for your participation.

### I. Your attitudes about and use of data

1. Current practices used to manage data in my field are holding back advances in knowledge. (circle one)
   Strongly disagree | disagree | undecided | agree | strongly agree
2. Information is being lost in my discipline by lack of attention to and effective strategies for preserving research data. (circle one)
   Strongly disagree | disagree | undecided | agree | strongly agree
3. In the course of my work, my research staff or I produce data frequently. (circle one)
   Strongly disagree | disagree | undecided | agree | strongly agree
4. My research staff or I obtain or contribute data to institutional, disciplinary, or community repositories as part of our work. (circle one)

Strongly disagree | disagree | undecided | agree | strongly agree

5. My research staff or I am aware of uses, other than end-stage publication of results, for the products of our analyzed or processed data. (circle one)
   Strongly disagree | disagree | undecided | agree | strongly agree
6. The data that my research staff or I predominately work with is produced by: (check all that apply)
   ____ controlled laboratory experiment
   ____ fieldwork observation
   ____ intermediate analysis
   ____ modeling or simulation
7. The size of the datafiles that my research staff or I predominately work with is: (check all that apply)
   ____ 1 megabyte (MB) or less
   ____ 1 to 100MB
   ____ 100MB to 1 gigabyte (GB)
   ____ greater than 1GB
8. Actions that my research staff or I frequently take when we work with our data are: (check all that apply)
   _____ data cleaning
   _____ data conversion
   _____ merging of datasets
   _____ data calculation
   _____ data visualization
   _____ none of the above
9. Please list the names of tools that help your research staff or you work with your data:
10. Please list some of the purposes your research staff or you have when you analyze or process your data:
11. Please list the names or web addresses of the institutional, disciplinary, or community data repositories your research staff or you use:
12. Please enter reactions, stories, or additional information prompted by the questions in this section regarding your attitudes about and use of data.

## II. Your management of data

13. I am satisfied relying on the durability and access commonly available computer storage media to determine the length of time the data produced by my research staff or myself is accessible. (circle one)
    Strongly disagree | disagree | undecided | agree | strongly agree
14. My research staff or I consistently enter and maintain information about the datasets we produce, such as the conditions, methods, and instruments used, which help us to later access and use the data. (circle one)
    Strongly disagree | disagree | undecided | agree | strongly agree
15. My research staff or I have used informational elements about datasets we've obtained from other researchers or central repositories to conduct your own research. (circle one)
    Strongly disagree | disagree | undecided | agree | strongly agree
16. The amount of effort my research staff or I currently spend managing our data, compared with other necessary research activities, is enough to maximize publication of our work results. (circle one)
    Strongly disagree | disagree | undecided | agree | strongly agree
17. The skills my research staff or I have to manage our data, compared with other necessary research skills, are enough to maximize publication of our work results. (circle one)
    Strongly disagree | disagree | undecided | agree | strongly agree
18. Sources that help my research staff or I decide what informational elements to apply or embed that describe my datasets are: (check all that apply)
    _____ my own planning
    _____ discussions in the lab or research group
    _____ examples of peer researchers
    _____ discipline-based requirements or standards
    _____ research center requirements or standards
    _____ information organization requirements or standards
    _____ government requirements or standards
19. The method(s) my research staff or I use to preserve and maintain my data for future access is: (check all that apply):

_____ Files managed in a PC

_____ CDs/DVDs, Tapes, or other removable media stored in the lab or office

_____ Networked file directories

_____ Content management systems

_____ Institutional, disciplinary, or community data repository

_____ Other

20. Please enter reactions, stories, or additional information prompted by the questions in this section regarding your management of data.

## III. Your professional status and affiliation

21. What is your present academic rank? (circle one)
    Professor | Associate Professor | Assistant Professor | Instructor | Other
22. What is your tenure status? (circle one)
    Tenured | On tenure track | Not on tenure track
23. Your current service in an administrative position: (circle one)
    Dean | Department Chair | Other | None
24. Do your interests lie primarily in teaching or research? (circle one)
    exclusively research | more research | more teaching | exclusively teaching
25. Please select the most appropriate field that describes both your background and current affiliation. (check one per column)

| Field | Highest Degree | Department of Appointment |
|---|---|---|
| Agriculture | | |
| Forestry | | |
| Bacteriology, Molecular Biology | | |
| Biochemistry | | |
| Biophysics | | |
| Botany | | |
| Environmental Science | | |
| Marine (life) Sciences | | |
| Zoology | | |
| General, Other Biological Sciences | | |
| Science Teaching | | |
| General, Other Education Fields | | |
| Aero-/Astronautical Engineering\ | | |
| Bioengineering | | |
| Chemical Engineering | | |

Table (Continued)

| Field | Highest Degree | Department of Appointment |
|---|---|---|
| Civil Engineering | | |
| Computer Engineering | | |
| Electrical Engineering | | |
| Environmentall Engineering | | |
| Industrial Engineering | | |
| Mechanical Engineering | | |
| General, Other Engineering Fields | | |
| Political Science, Government | | |
| Mathematics and/or Statistics | | |
| Astronomy | | |
| Atmospheric Sciences | | |
| Chemistry | | |
| Earth Sciences | | |
| Geography | | |
| Oceanography | | |
| Physics | | |
| General, Other Physical Sciences | | |
| Archaeology | | |
| Cognitive Sciences | | |
| Experimental Psychology | | |
| Social Psychology | | |
| General, Other Psychology | | |
| Economics | | |
| Sociology | | |
| General, Other Social Sciences | | |
| Computer Science | | |
| Data Processing, Computer Prog. | | |
| Communications | | |
| Library and Information Science | | |
| Information Studies | | |
| All Other Fields | | |

# Appendix B

## Topics Covered in the Science Data Management Course

| Date | Topics |
|---|---|
| Week 1 | Introduction to the course<br>Science data life cycle and basic concepts |
| Week 2 | Fundamentals about data:<br>forms, types, levels of processing<br>Data structures and models: physical data, model data<br>Data formats: data format standards<br>Representation of data<br>Communication of data |
| Week 3 | Describing datasets (1)<br>Metadata as solution to science data management<br>Metadata types<br>Metadata standards for scientific datasets |
| Week 4 | Describing datasets (2)<br>Examples of metadata records for science datasets<br>Principles and requirements<br>Adoption of metadata standards |
| Week 5 | Data provenance<br>Understanding science workflows<br>What is data provenance?<br>Provenance metadata<br>Data provenance and curation |
| Week 6 | Relational databases<br>Data attributes<br>Data relationships<br>Databases<br>Example data sets in relational databases |
| Week 7 | Managing data with repositories<br>Data management tasks<br>Data curation tasks<br>User requirements for data repositories<br>Technical requirements for data repositories |
| Week 9 | Challenging issues in data repository services<br>Interoperability<br>Discovery and search<br>Ownership and access<br>Security<br>Evaluation |
| Week 9 | Data use scenario analysis<br>Guest speaker: a bioinformatics scientist |
| Week 10 | Data management scenario analysis<br>Guest speaker: a biophysicist |

Table (Continued)

| Date | Topics |
|---|---|
| Week 11 | Developing data management project (1):<br>Data set characteristic analysis<br>Needs assessment<br>User roles (researchers, lab staff, etc.)<br>Planning<br>Goals and objectives<br>Procedures<br>Policy development<br>Quality control |
| Week 12 | Developing data management project (2):<br>Data organization and preservation<br>Metadata issues<br>Long-term preservation of data<br>Enabling technologies:<br>for organizing and managing data<br>for storing and retrieving data<br>for using data |
| Week 13 | Data quality, discovery, and publishing<br>Read metadata description<br>Quality criteria<br>Data repositories and discovery<br>Directory services<br>Data publishing and sharing |
| Week 14 | Data analysis<br>Data mining<br>Data meshing-up<br>Data presentation<br>Visualization: tools<br>Formatting for Publication<br>Using data<br>Ethics<br>Citing datasets |
| Week 15 | Project presentations and discussions<br>Wrap-up<br>Post-course assessment survey |