2011

Determining Data Information Literacy Needs: A Study of Students and Research Faculty

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

Michael Fosmire
Purdue University, fosmire@purdue.edu

Chris Miller
Purdue University, ccmiller@purdue.edu

Megan R. Sapp Nelson
Purdue University, msn@purdue.edu

Determining Data Information Literacy Needs: A Study of Students and Research Faculty

Jacob Carlson, Michael Fosmire, C.C. Miller, and Megan Sapp Nelson

jrcarlso@purdue.edu; Fosmire@purdue.edu; ccmiller@purdue.edu; mrsapp@purdue.edu

Purdue University Libraries
504 West State Street
West Lafayette, IN 47907-2058
Phone: (765) 494-2859; Fax (765) 494-0706

Determining Data Information Literacy Needs: A Study of Students and Research Faculty

Abstract

Researchers increasingly need to integrate the disposition, management and curation of their data into their current workflows. However, it is not yet clear to what extent faculty and students are sufficiently prepared to take on these responsibilities. This paper articulates the need for a data information literacy program (DIL) to prepare students to engage in such an "e-research" environment. Assessments of faculty interviews and student performance in a geoinformatics course provide complementary sources of information, which are then filtered through the perspective of ACRL's information literacy competency standards to produce a draft set of outcomes for a data information literacy program.

Introduction

The nature and practice of research and scholarship is undergoing dramatic change with the advent of ready access to high bandwidth networks, the capacity to store massive amounts of data, and a robust and growing suite of advanced informational and computational data analysis and visualization tools. The practice of technology-driven research, known as e-science, or more broadly as e-research, has had a transformative effect in the science and engineering fields. E-research applications are growing within the humanities and social science disciplines as well, where e-research is poised to have similar dramatic effects on the nature and practice of research.

The complexity and scale of e-research in turn requires an evolution of traditional models of scholarly communication, library services, and in many cases the fundamental role of librarians themselves. In response, librarians have started to initiate discussions and projects to situate librarians in those areas of e-research most in need of library scientist expertise.[i] In light of the new expectation of the National Science Foundation that every grant proposal needs to have a data management plan,[ii] libraries are beginning conversations in their universities to negotiate a role in the management of research outputs.

Data also provide the opportunity for an evolution of library services in the area of instruction. Most academic libraries already offer information literacy courses and programs as a part of their services. Extending these information literacy efforts to include programs on data management and curation may provide a logical entry point into increasing libraries' role in supporting e-research. A successful education program, however, must be based on a firm understanding of current practice and standards as well as the needs of the target audience. There is a lack of research on the needs of both the research practitioners and the students grappling with these issues in the classroom and in the laboratory. The authors attempt to address this knowledge gap by gathering data from interviews with faculty researchers and from information gathered from the authors' own geoinformatics course. Based on these data, the authors propose a model set of outcomes for data information literacy (DIL).

Background

E-Research and Implications for Libraries

E-research has had a tremendous impact on a number of fields, increasing the capabilities of researchers to ask new questions and reduce the barriers of time and geography to form new collaborations. In astronomy for example, the National Virtual Observatory (NVO) makes it possible for anyone from professional astronomers to the general public to find, retrieve and analyze vast quantities of data collected from telescopes all over the world.[iii] For scholars of literature, the Metadata Offers New Knowledge (MONK) project not only provides a collection of digitized texts, but also tools that can be used to apply data mining, visualization and other techniques towards the discovery of new patterns and insights.[iv] It should be no surprise, of course, that such projects simultaneously produce and feed upon large amounts of data. The capture, dissemination, stewardship and preservation of digital data have therefore been identified as critical issues in the development and sustainability of e-research.

Funding organizations and professional societies have identified a need for the development of educational initiatives to support the development of a workforce capable of supporting the development of e-research initiatives. The National Science Foundation (NSF) first described the connection between e-research and education. The NSF's Atkins report, released in 2003, highlighted the need for highly coordinated, large-scale investments in several areas, including developing skilled personnel and facilities needed to provide operational support and services.[v] In 2005 the National Science Board produced a report that articulated existing and needed roles and responsibilities required for stewarding data collections, followed by a series of recommendations for technical, financial and policy strategies to guide the continued development and use of data collections.[vi] The American Council of Learned Societies issued a report in 2006 calling for similar attention and investments in developing infrastructure and services for e-research in the Humanities fields.[vii] More recently, the National Academy of Sciences issued a report advocating the stewardship of research data in ways that ensure research integrity and data accessibility. The recommendations issued in the report include the creation of systems for the documentation and peer review of data, data management training for all researchers, and the development of standards and policies regarding the dissemination and management of data.[viii]

While the rich, collaborative, and challenging paradigm of e-research promises to produce important, even priceless cultural and scientific data, librarians are still determining their role in the curation, preservation, and dissemination of these assets. In examining how e-research may affect libraries, Hey and Hey argue that e-research "is intended to empower scientists to do their research in faster, better and different ways."[ix] They particularly emphasize that information and social technologies make e-research a more communal and participatory exercise, one that will see scientists, IT staff, and librarians working more closely together than they have previously.[x] A particular challenge looming

with the rise of e-research is the “data deluge”, the need to store, describe, organize, track, preserve, and interoperate data being generated by a multitude of researchers in order to make the data accessible and useable by others for the long-term. The sheer quantity of data being generated and our current lack of tools, infrastructure, standardized processes, shared workflows, and personnel who are skilled in managing and curating these data pose a real threat to the continued development of e-research.

Gold provides an outline of the issues and opportunities for librarians in e-science. Starting from the familiar ground of GIS, bioinformatics, and social science data, Gold argues that librarians working in e-science will be busy developing relationships – both upstream and downstream of data generation, and the effort may be “both revitalizing and transformative for librarianship.”[xi] Similarly, the *Agenda for Developing E-Science in Research Libraries* outlines five main outcomes that focus on capacity building and service development in libraries for supporting e-science.[xii] Walters further asserts that libraries taking "entrepreneurial steps" toward becoming data curation centers are on the right track, reasoning that "a profound role for the university research library in research data curation is possible. If the role is not developed, then a significant opportunity and responsibility to care for unique research information is being lost."[xiii] In other words, the community seems reasonably sure that supporting e-research is not so novel that it falls outside of the mission and founding principles under which libraries have operated for decades.

Educational Preparation for E-Research

Ogburn predicts that e-science will quite certainly fail if future generations of scholars are not savvy with *both* the consumption and production of data and tools. “To prepare the next generation of scholars the knowledge and skills for managing data should become part of an education process that includes opportunities for students to contribute to the creation and the preservation of research in their fields.”[xiv] Here is data information literacy from thirty thousand feet, it is not simply enough to teach students about handling incoming data, they must also know, and practice, how to develop and manage their own data with an eye toward the next scientist down the line. The Association of Research Libraries reported to the NSF in 2006 that because “many scientists continue to use traditional approaches to data, i.e., developing custom datasets for their own use with little attention to long-term reuse, dissemination, and curation, a change of behavior is in order…[This change] will require a range of efforts, including investment in approaches to make data documentation, sharing, and preservation easier, establishment of an infrastructure to accept and assume responsibility for data...and, perhaps most important of all, concerted efforts to educate current and future scientists to adopt better practices.”[xv]

The inspiration for the authors' own work on instructional components to e-science comes from the National Science Foundation's *Cyberinfrastructure Vision of 21st Century Discovery*, in which the dramatic rhetoric of revolution and recreation does indeed trickle down to education:

> " … Curricula must also be reinvented to exploit emerging cyberinfrastructure capabilities. The full engagement of students is vitally important since they are in a

> special position to inspire future students with the excitement and understanding of cyberinfrastructure-enabled scientific inquiry and learning. Ongoing attention must be paid to the education of the professionals who will support, deploy, develop, and design current and emerging cyberinfrastructure."[xvi]

Although many have articulated the need for educating a workforce that understands the importance of managing and curating data in ways that support its broad dissemination, use by others, and preservation beyond the life of its original research project, there has been very little examination of what such a program would contain. We believe that librarians have a role in developing these education programs and will need to actively engage in these discussions.

Gabridge notes that institutions experience "a constantly revolving community of students who arrive with…uneven skills in data management….Librarian subject liaisons already teach students how to be self-sufficient, independent information consumers. This role can be easily extended to include instruction on data management and planning."[xvii] With the respectful elision of "easily," we argue in the remainder of this paper that there are indeed gaps in the knowledge of current e-researching faculty and students (both as producers and consumers of data) that librarians may help fill by developing a data information literacy (DIL) curriculum.

Environmental Scan of Related Literacies

For the sake of clarity, it is important to distinguish DIL from prior literacies such as data literacy, statistical literacy, and information literacy. Typically, data literacy involves understanding what data mean, including how to read graphs and charts appropriately, draw correct conclusions from data, and recognize when data are being used in misleading or inappropriate ways. Statistical literacy has been defined as "the ability to read and interpret summary statistics in the everyday media: in graphs, tables, statements, surveys and studies."[xviii] Schield finds common ground in data, statistical, and information literacy, stating that information literate students must be able to "think critically about concepts, claims, and arguments: to read, interpret and evaluate information." Furthermore, statistically literate students must be able to "think critically about basic descriptive statistics, analyzing, interpreting and evaluating statistics as evidence." Data Literate students must "be able to access, assess, manipulate, summarize, and present data." In this way, Schield creates a hierarchy of critical thinking skills: data literacy is a requisite for statistical literacy, and statistical literacy required, in turn, for information literacy.[xix] Stephenson and Caravello[xx] extol the importance of data and statistical literacies as components of information literacy in the social sciences, arguing that the ability to evaluate information essentially requires that one understand the data and statistics used in an information resource.

Qin and D'Ignazio have developed a model to address the production aspect of data management, called Science Data Literacy. SDL refers to "the ability to understand, use, and manage science data," and an SDL education "serves two different, though related, purposes: one is for students to become e-science data literate so that they can be effective science workers, and the other is for students to

become e-science data management professionals. Although there are similarities in information literacy and digital literacy, science data literacy specifically focuses less on literature-based attributes and more on functional ability in data collection, processing, management, evaluation, and use." [xxi]

Whereas definitions of data, statistical, and information literacy focus on the consumption and analysis of information, the authors believe that the production of information is an important component often overlooked in literacy instruction. E-research is, by definition, a social process, and contributing to – not just extracting from – the community's knowledgebase is crucial. Data information literacy, then, merges the concepts of researcher-as-producer and researcher-as-consumer of data products. As such it builds upon and reintegrates data, statistical, information and science data literacy into an emerging skill set.

Prior Instructional efforts in data information literacy

Although libraries have not uniformly determined that they should provide instructional support for data information literacy competencies, several libraries have already developed programs or prototypes to address those needs. The Massachusetts Institute of Technology Libraries created a robust "Manage Your Data" subject guide/tutorial, supplemented by seminars such as Managing Research Data 101. Both resources include data planning checklists that include the following topics:

- Documentation and metadata
- Security and backups
- Directory structures and naming conventions
- Data sharing and citation
- Data integration
- Good file formats for long-term access
- Best practices for data retention and archiving[xxii]

The University of Virginia Libraries created the Scholar's Lab and Research Computing Lab in 2007, which merged into a single entity in 2010. These projects, collaborative ventures between information technology and libraries departments, created a new service model that includes traditional roles for IT (software support and training) and librarians (subject knowledge and departmental interactions) as well as services that bridge those disciplines such as data management and analysis, computational software support, and knowledge of emerging technologies. As librarians from UVa explain, "we chose to promote the service areas of software support, current awareness, data, collaboration, and research communication…. Collectively, we view these as being supportive pieces to the entire research lifecycle, rather than just a single point."[xxiii] While the University of Virginia model focused primarily on reference and project based services, the Scholar's Lab also provided workshops and seminars on special topics in data management such as GIS, web application development, and text digitization.

The Science Data Literacy Project at Syracuse University has developed a program "to train students with the knowledge and skills in collecting, processing, managing, evaluating, and using data for scientific inquiry."[xxiv] As part of the project, Jian Qin developed a credit-bearing course, "Science Data Management," covering the fundamentals of scientific data and its description, manipulation, visualization, and curation. Project SDL makes its syllabus for the course, with lecture notes, available online.[xxv]

The Purdue University Libraries have been active in this area as well. Two of the authors developed a geoinformatics course with a faculty member in the Earth and Atmospheric Sciences department, which has been offered in Spring of 2008 and Spring of 2010.[xxvi] The instructors designed Geoinformatics for beginning graduate and advanced undergraduate students. The course provides a holistic approach to GIS and spatial data, not focusing specifically on GIS as desktop analysis but rather encompassing the fuller cycle of data, from discovery and acquisition to conversion and manipulation, analysis, and finally, visualization, metadata, and re-sharing.

The syllabus for the 2008 and 2010 courses are available online.[xxvii] More information about the course and student learning will be presented later in this article.

Assessments of Faculty and Student Needs in Data Information Literacy

Like e-research, data information literacy is not new in and of itself, but rather compiles expertise and portions of existing research methods, information and other literacies, and computing curricula to offer more holistic, communal, and participatory perspectives and techniques for future e-researchers. Just as e-research encourages researchers from a variety of disciplines to collaborate to advance scientific knowledge, disciplinary and library faculty must work together to determine the skill sets that a data literate student should demonstrate and to develop best practices for imparting those skills to the students. Both faculty members and students have perspectives on what the necessary data management skill sets in their fields should contain. These perspectives on data management are grounded in their real-world perceptions and practices and a first-hand knowledge of how research is conducted in their respective disciplines. Any attempt to define a data information literacy program must be aligned with current disciplinary practices and cultures if it is to be relevant to and accepted by its intended audience(s). The authors compiled the perspectives of both faculty and students from two different research projects, one based upon interviews with faculty members and the other upon surveys of students and an analysis of their coursework. In the next two sections, the authors report on the DIL priorities articulated by both faculty and students as discovered through our assessments.

Faculty Assessment

Faculty Assessment - Methodology

In the Fall of 2007, the Purdue University Libraries and the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign (UIUC) received funding from the Institute of Museum and Library Services (IMLS) to conduct research on the willingness of research faculty to share their data with others (including the conditions necessary for data sharing to take place) and to investigate possible roles for librarians in facilitating data sharing and curation activities.

The investigators interviewed participating faculty at Purdue and UIUC, focusing on three broad areas: the nature and lifecycle of one of the data sets generated by the researcher; his or her data management practices; and their needs for making their data available to others and curating their data for long-term access. These interviews resulted in the creation of "data curation profiles," each of which summarized the information gathered from the interview under a common framework that enabled comparisons to be made among the researchers' responses.[xxviii]

The first round of interviews for this project took place at Purdue and UIUC in the summer and early fall of 2008. Faculty participants were not randomly selected, but recruited from a broad selection of departments in the sciences and engineering, based on prior relationships with project personnel or liaison librarians. The semi-structured interviews asked deliberately broad and open-ended questions to allow participants to control the direction of the discussion and identify the most important issues they found related to sharing and curating their data. The investigators then extracted common themes from the transcripts using a grounded-theory approach.

One of the common themes emerging from the interviews concerned the skills, knowledge, and training needed by graduate students to effectively manage and curate research data. Graduate students actively generate and curate data in support of their own research. Many also oversee the management of data generated by the entire research group. A few of the faculty also noted that their graduate students had been asked to share their data with individuals not affiliated with the research and therefore had to consider similar issues of whether or not to share and what conditions to place on sharing. Typically, faculty determined graduate students were unprepared to manage or curate the data effectively, while acknowledging that although this was an area of concern for them, they often could not provide adequate guidance or instruction because it was not an area that they knew well or fully understood.

The investigators conducted a second round of interviews in the spring of 2009 to gather more detailed and quantifiable information from faculty and address any gaps in information from the first interview. In this second round, investigators asked the faculty participants at Purdue whether there was a need for a data management and curation training program for graduate students, and what such an educational program should contain. Responses from these second interviews were then coded and analyzed in conjunction with the relevant information from the first interviews. A total of 19 faculty from both schools completed both interviews.

Faculty Assessment – Results

Overview

Generally, faculty in this study expected their graduate students to be able to carry out data management and handling activities. However, the extent of data management responsibilities varied among the faculty interviewed, with some taking an active, hands-on role in managing their data with minimal student involvement, while others delegated most data management tasks to their students. Typical responsibilities of graduate students included processing or cleaning the data to enable use or analysis, assuring quality of the data, compiling data from different sources, and organizing the data in ways that it could be accessed and used by project personnel. In addition, faculty often considered data management duties as distinct from other research responsibilities. In the words of one professor,

> "[I ask the graduate student] basically to put the data together…[to] look at it, clean it up, talk to me if you don't know how to clean it up, put it in a form that we can do these… calculations, and depending on who they are and how good they are, they may actually do some of those calculations…" (Agronomist #3)

Analysis of the interviews revealed that the training graduate students received and the methods through which this training was delivered varied widely. Some of the researchers taught their graduate students data management tasks, such as how to develop and assign metadata to the data files. Other researchers reported that their graduate students had not received much if any formal training in data management and were left to figure things out on their own. Reflecting on current practice, one researcher identified major problems with the way students learn data information literacy skills.

> "…The way we [teach these skills], is generally speaking, we just say, 'Well, go learn it' and [the graduate students] just figure it out. And the problem with just figuring it out is every grad student is going to do it a little differently. You know, there's no standard so interoperability or even just trying to preserve it long term or even knowing how to hand it off… It's difficult to do that once the student expert that made those programs in that data has left." (Electrical Engineer)

Given the variance in the range of responsibilities and training in data management received by graduate students, it is not surprising that faculty in this study presented a mixed picture in assessing the work of their students in this area. Several faculty expressed frustration with their inability to understand or make use of the data their students had been working on, especially after they graduate.

> "[I]f I'm lucky, I get one student that starts before the next one leaves, and that's the handoff, here's the tribal knowledge, here's the data. If I'm unlucky, one of them graduates in May, and the next one doesn't come till August, and then I'm slogging through CDs and DVDs… trying to get less-than-perfectly archived data into a usable format… sometimes it takes a week of goin' through… people are inventing filename schemes, and everybody invents their filename scheme differently, and they tab it differently, and, some of them put their date and time in a nice Microsoft format, other people put it in milliseconds, sometimes people import text fields…" (Civil Engineer)

Other comments provided a positive statement of individual students' skills, which they generally acquired without formal training. One researcher detailed an exceptional graduate student's deliverables:

> "He left me … a complete description of every file, and how it was organized… by folders, by file names…printed out, all nice…here's the master key for what all these files are, and what I used them for, and where…everything from the SAS program that went with the analysis, so…there are lots of things that could be linked to things, so if you do an analysis, you could link the SAS program, so that if somebody wanted to run the same analysis but do something different with…the data, the SAS program that you used is written and coded and…it's right there…so …if you have SAS you just open the run program, and play with the data whatever way you want, and re-run the data." (Agronomist #2)

The overwhelming majority of researchers in this study felt that some form of data information literacy education was needed for their students. However, even in stating a need for such a program, several respondents expressed an uncertainty or a reluctance to teach data management skills to their students themselves. Some faculty expressed a concern about getting too involved in telling students what to do in what should be the student's own work, or in making their work more difficult by introducing new software or formats to work with.

> "[Decisions regarding data management tools] just kind of [turn] out to be the one the students do. How much do I micromanage on that? [It] gets to be difficult." (Civil Engineer)

Furthermore, although faculty identified the lack of data management skills in their graduate students as a strong concern and described broad themes that should be addressed, they often could not articulate precisely what skills should be taught to remedy the situation.

> "Interviewer: Is there a need for education in data management or curation for graduate students…?
>
> Faculty: Absolutely, God yes… I mean we're…We have the ability to accumulate huge datasets now especially with the new tools that we have…
>
> Interviewer: So, what would that education program look like, what would it consist of? What kind of things would be taught?
>
> Faculty: Um, I would say, um, and I don't really know actually, just how to do you manage data? I mean, where do you put it? Um, how secret does it need to be? Or you know, confidentiality things, ethics, probably um…I'm just throwing things out because I hadn't really thought that out very well."
>
> (Soil Scientist)

After coding and analysis, several major themes emerged from the faculty's observations of graduate students' deficiencies in data management. These themes are metadata, standardizing documentation

processes, maintaining relationships among data, ethics, quality assurance, basic database skills, and preservation.

Metadata

An understanding of metadata and how to apply it were frequently mentioned as areas of need, although "metadata" as a term was not typically used. More often, researchers said their students needed to know how to annotate and describe data. In most cases, references to "annotations" included both a need to provide information about a data file as well as information about individual components of the data (such as a cell in a spreadsheet). The main reasons for providing metadata include the assurance that data can be understood by others (both within the lab and by external audiences), enabling its continued usability over time, and fostering use of the data beyond its original purpose. As one professor stated,

> "[Students need to be taught]… how to organize the data and then… how would you annotate things in a way that would make [the data] useable not only to others but to yourself 6 months from now. You know, we try to tell students… to take notes and that's why we have lots and lots of notes because you might want to go back to that 6 months from now, you don't quite remember, and so you have notes but now they should also be doing that somehow with the data once it gets in the set." (Agronomist #3)

Researchers also expressed the need to apply and conform to metadata standards. One researcher stated that not only must students be taught "how to approach the idea of metadata," but also develop an awareness of standardized disciplinary ontologies and how to apply them to their own work. "…a student knows what he's done, but sometimes he hasn't written it down as detailed as you would like and you're trying to read someone else's notebook and [figure out] what did he mean because it is not in an ontology." (Biologist - U1B1B1, 18:10)

Standardizing Documentation Processes

This rather broad theme includes both high-level organization as well as more specific, local needs.

> "…I think [students] should learn some good data management hygiene, I mean most of them don't even know how to organize their files in a logical way, so they don't [even] start with a method… So organization is a pretty big deal in general." (Agronomist #2)

Researchers frequently reported a need for students to be able to organize data by documenting it in a systematic and logical fashion. Explanations given for the need for rich documentation often extended beyond the immediate needs of the researcher's lab and included such high-level needs as enabling the sharing of data outside the research team, submission to repositories, re-use by external audiences, and preservation beyond the research lifecycle. At the more local level, this category

addresses folder and file naming conventions, data sharing amongst the lab/project team(s), and defining staff responsibilities for managing data, communication, and workflow.

Researchers expect their graduate students to share responsibility for documenting the lab or project's data, as well as the student's own interactions with it. Documenting data focuses on what needs to be recorded and provided while generating, processing, analyzing, and/or publishing the data in order to later validate and verify it. This may include such tasks as generating and maintaining data dictionaries, glossaries or definitions of variables, maintaining lab notebooks or their equivalent, and capturing the provenance of the data. Overall, researchers expressed that students' documentation needs to stand the test of time. "I tell [the students] that their documentation should be good enough that when I go look at it ten years from now I still understand what you did." (Biochemist)

Researchers in this study acknowledged the problem of data documentation, not only for their students but for themselves as well. One interviewee describes the situation as follows:

> "Well, every student…[is] continuing to evolve and create stuff and the longer you're here the more they accumulate…facts – some of the stuff gets written down, some of it doesn't…I wish we all were perfect documenters [but that's not the case]" (Civil Engineer)"

Difficulties in documenting data contribute to a larger concern: the lack of standardization and consistency in how the data are organized. Faculty repeatedly mentioned that every student employs different methods of documenting their data. The lack of standardized and shared data management protocols and practices across a research group often leads to a "tower of Babel" situation, leading to difficulties in correlating and relating one data file with another or with the data collection as a whole.

> "…different students have done different things. I have a way of, I keep my statistical programs in one file, and my data files for running in another file. Raw files are kept someplace else and labeled as raw. But different students do different things and at different times. And then they figure out that it would be better to do this and they suddenly change… paying attention to data protocol is like an additional level, it's relatively new and we are not very good at it." (Agronomist #2)

The inevitable turnover of students over time exacerbates this problem. Although most of the researchers in this study require their students to document their work with the data, actual documentation practices followed by the students varied from one to the next. Moreover, they often did not provide complete or detailed enough documentation to enable others to understand their work.

> "So there was no very formal way of doing this …you'd have one graduate student and you would tell them how to do it and they would do it and then they would leave… I've had three graduate students who've worked on various parts of this. And they've shared their…data records, but it's all fairly informal… [I]t involves here's my data…and let me sit down with you and explain to you what this means, and this means, and this means." (Agronomist #2)

Several researchers suggested creating a standard operating procedure for data formatting and management. One faculty member noted that they created standard operating procedures for most equipment and procedures in the lab and proposed that a similar standard operating procedure be developed for handling and managing her data. When asked to describe an ideal situation for organizing data, several of the faculty members noted the need for students to develop and use a standardized set of best practices. As one faculty stated:

> "We have [the students] document what they do, but we don't have… one of the things that would be nice would be sort of a best practice, you know. You use these naming conventions and we sort of standardize that across students, but we haven't managed to do that." (Electrical Engineer).

Maintaining Relationships among Data—Master Files and Versioning

Many interviewees described the challenge of relating data files to each other. This includes issues related to taking data generated at one site at one time or for a particular purpose and defining the relationships between those files and/or enabling that data to be integrated to create a new data file or data set. This category also includes the converse action, generating a subset of the data from a larger data set or file.

Several researchers specifically mentioned the need for the creation of an official record of the data (a "master file") to ensure the authority and integrity of this record compared to the working copies of data sets or files created and used for specific purposes by subsets of lab/project personnel. One researcher described such a file he charged a student with creating:

> "…the last PhD student I had…, I had her bring [a variety of files] together because…she started asking questions about relationships across time, space. …I had her marry [the data] together…so that we had a good, a complete assessment of what's going on, and…that's a very precious file I call a master file…That's the archive, that's the record." (Agronomist #1)

Many researchers additionally desired that the master file bring a number of disparate files together into a searchable database that engenders question development and helps assure quality control for research. A lack of standardization in data management practices, a high learning curve, and a perceived lack of support for the advanced database utilities and programs required to create such files hinder their ability to achieve those goals.

Researchers also expressed the need to balance the requirements for a particular research project with those for making the data accessible and useful to the larger research community. One researcher articulated this as follows:

> "…Typically, what's happened is that you have a student, you tell them where the data are, you tell them what to look for…and they go and they collect the various Excel spreadsheets and merge them into a file… It's never all that useful for the next question because it's very specific to the question that they're asking." (Agronomist #2)

This focus on the specific research needs of the student (or the faculty sponsor in some cases) often led to situations in which the faculty member could not retrace the steps taken in processing the data and relate the graduate student's work back to the larger data set to which it belonged.

Akin to these issues of compiling or merging data, researchers frequently brought up versioning as an often neglected but very important concept for students to learn. When asked about skills that her students need to know one faculty member replied:

> "…how do you keep track of multiple versions of files and remember which things are which and annotate it so you know which things are which." (Agronomist #3)

In this study, researchers clearly reported the importance of maintaining documentation of different versions of their data. They wanted to know which data files were used for what analysis, what file contained the current version being used by the research group, and how these versions differed from each other. However, several faculty members admitted that they themselves had a difficult time in maintaining adequate documentation and struggled to consistently generate the needed documentation in a timely manner.

Ethics

Faculty members in this study identified "data ethics" as another area where most students need assistance. Data ethics includes intellectual property rights and ownership of data, issues of confidentiality/privacy and human subjects, implications and obligations of sharing (or not sharing) your data with others (including open access), and assigning attribution and gaining recognition of one's work. Although faculty clearly stated ethics as a needed area of instruction for their graduate students to understand and abide by, they generally did not provide much description as to what the curriculum of such an ethics program would include. In one case, the professor tied ethics to an understanding of ownership of data.

> "A portion of [a student's education] is actually data ownership, data ethics, because oftentimes students are unclear about you know, it's my data. And they leave with their data. No, it's actually not your data…So, something that rapidly brought them up to speed on some basic how-tos, as well as the ethical…perspective…that they would need to have when they become the person who's in charge [of a research lab]." (Agronomist #2)

Basic Database Skills

Several researchers expressed the expectation that students be able to understand and develop relational databases and use database tools effectively. Frequently, students' lack of basic understanding of database development and usage frustrated the interviewees. However, the expectations of student skills differed among the researchers. One professor reflected:"…[T]he students coming in…don't know anything about getting the database, the data in first normal form,

second normal form, third normal form, primary keys, all that sort of stuff. I could see [the libraries and departments] working collaboratively, saying okay…here's the database warehouse or engine that we provide, here are the skills that we'll help you set up these tables, we'll set up some ODBC links or some links so that you can pull reports out…" (Electrical Engineer)

Another professor acknowledged that not all students necessarily needed training in SQL, but most did need some basic understanding of relational databases, normalization of data, database tools and documentation techniques. (Civil Engineer)

Quality Assurance

Researchers expect their graduate students to review or check their data and evaluate its quality.

> "[Graduate students should be taught] how to organize data and something about how to check the data. What I find with a lot of students, I mean, even in my classes…, they don't ever look at the data to see if it looks reasonable. Something about looking and checking or cleaning up...and getting them to think about…Just because this number comes off a machine, doesn't mean it's right."(Agronomist #3)

> "[The graduate student is] looking all the time to see…number one, is the data good? [T]he grad students usually checks that right off the bat on the day he's down there taking the data. If it isn't running right he tries to restart it, he tries to tweak it, twiddle it, fiddle it, and that sort of thing. …[W]e will occasionally find that some site is kind of noisy and it can, if it's not too noisy it's part of the…interesting part of the problem, you know, how robust are our algorithms to these noisy sites? On the other hand, if this site is just really bad or offline then of course we can't get any data, at all." (Electrical Engineer)

Interviewees mentioned the difficulty knowing exactly what their students had done to compile and analyze the data, thereby making the data of unknown provenance. One professor stated that she could not always understand the work done by her students in reviewing her data:

> Faculty: …working in this area, is a kind of unknown territory, and so if you give it to students or post-docs it is very hard to troubleshoot, so that's the reason I have to do lots of it myself.
>
> Interview: What do you have to do when you have to troubleshoot? What kind of things pop up?
>
> Faculty: … it gets kind of bizarre results, when the students do it, then you don't know whether it is real, or if it was mistake. So, when I do it, then know what I did. If my student does it, then I don't know what I did. So it kind of takes more time to know what happened. (Biochemist)

Quality assurance is in some ways a blend of technical skills (familiarity with equipment), disciplinary knowledge (whether the result is even theoretically possible), and a metacognitive process that requires synthesis on the part of the students. Primarily, quality assurance is the ability to recognize a

pattern or consistency in the data. Quality assurance is facilitated or disrupted by the quality of documentation (annotation/metadata) produced, and the organizational schema, or lack thereof, of a given data set.

Preservation

Researchers expect their students to know how to preserve their data and document the processing of the data.

> "[Graduate students should know] how you preserve things for a long time. How [to] preserve the record of your scientific work… [They should know] a way to standardize [data] somehow so that the faculty or anyone interested…could go back in it, [and] check the results. Just do the kind of thing that good science would expect people to be able do to verify that things are correct." (Electrical Engineer)

Much like the discussion of metadata, faculty members generally understood the term "preservation" in a broad and loose sense of the word, often conflating it with the simple backing up of files. They were unaware of or unacculturated to "preservation" from a library perspective, instead focusing much more on the immediate issues and procedures surrounding backing up their data.

Although researchers recognized the need for backups, the methods and timing of performing the backups differed considerably among research groups. Some, having learned the hard way through lab disasters, kept geographically dispersed backups. (Agronomist #1) Others relied largely on graduate students to create backups on departmental servers (Agronomist #2; Electrical Engineer). Still others had no real-time backup system in place (Biologist). A common problem expressed with backups was tracking versioning.

> "I back up, I've got a, like an extra hard drive that I back up on… And then I try to make sure that my student has a copy, and so, it's a little bit difficult to have the project sort of at a final stage because we both have different copies of different data, and there's also a lot of information that we save, but we're not sure we're ever gonna use it, so it's, it's a, really a mess. Because these big data sets, we just don't quite know how to deal with them…" (Soil Scientist)

Faculty Assessment – Concluding Remarks

The design of any data information literacy (DIL) program requires an understanding of the real-world needs of research groups, where students labor in the trenches and the research either progresses or is impeded by their ability to handle data in the ways described here. As such, the faculty supervisors are privy to, and no doubt acutely aware of, the defects in their students' abilities to properly care for their research input and output. The interviews analyzed for this study provide a window into the

ground-level interaction with data and in fact become a magnifying glass through which we can spot the deficiencies and gaps in knowledge that a DIL curriculum might target.

We would be remiss, however, to not account for the "ignorance loop" in faculty responses, as these interviews also expose faculty interaction with data. Many faculty admitted or otherwise revealed that they themselves lack the expertise or experience with data management, even as they critiqued their students' abilities. We must assume their critiques of their students (and their own) facility with any or all aspects of data management may be somewhat shallow. In other words, they may not know what they don't know about data management and curation. Therefore, a program based entirely on faculty self report risks incompleteness and other viewpoints on what should constitute the objectives for DIL must be taken into account.

As a complement, then, the next section will draw conclusions that help to complete our DIL core objectives from a direct source, a course taught at Purdue University that broached some of these exact topics, including data source evaluation, metadata, databases, preservation, and sharing.

This course allowed us to examine the DIL of students directly and learn from first-hand observation. Because we gained insight into what the students do not know, our own evaluation of student performance in a (classroom-simulated) research environment can serve as an important second front in developing a richer and more comprehensive list of core data information literacy objectives.

Student Assessment

Student Assessment – Methodology

Enrollees in the 2008 and 2010 offerings of Geoinformatics provided the sample population for our student assessment. The combined number of students enrolled totaled 27 students, 12 in 2008 and 15 in 2010. Most of these were Earth & Atmospheric Sciences students, but other departments represented in this course included Civil Engineering, Agricultural & Biological Engineering, and Forestry & Natural Resources. In 2008, the core course content revolved around a ‘whodunit’ conceit, with students needing to track down, over the course of several laboratory exercises, the location of a fictitious chemical spill by gathering data (both spill data and underlying geology) and using various geospatial analysis and visualization techniques. Semester projects, mostly based on the students’ own research projects, provided the rest of the context for learning data information literacy skills. The 2010 course dropped the whodunit mechanism in order to shift more attention toward a longer, more involved semester project.

In order to improve and tailor the course, the authors used several methods to probe students’ interests, perceived needs, and their abilities to carry out data management tasks. Among these were a pre-course assessment to inventory the students' technology and information skills and a post-course survey to determine their perceptions of how important different topics were to their research. The

instructors also analyzed student semester projects to determine how well they demonstrated mastery of data information literacy skills.

The pre-course survey was administered in both offerings of Geoinformatics. It contained short-answer questions, mainly probing their background in databases, GIS, and programming, such as "What computer programming languages do you know (for example, Fortran, C)?" and "What geospatial software do you use?" The instructors then tailored the course content to address the ability levels of the students. The post-course survey was given only to students in the 2008 offering of the course. For each course topic, students rated, on a five-point Likert scale, the lectures, the lab, and the importance of the topic to the course and to their own research and they also recommend improvements to the course's labs.

These instruments probed students' attitudes toward various topics related to data information literacy. However, there were disconnects between student perceptions and their performance. As Grimes and Boening, among others, have observed, novices tend to overstate their expertise, in large part because they don't know what they don't know.[xxix] To provide a check of the degree to which students actually demonstrated data information literacy skills, the instructors analyzed student semester projects, which required students to identify a problem or research question with geospatial components and use the skills and techniques discussed in class to advance that research and present the results of their work. The project required both the acquisition of original data and the use of external, 'published' data, involved some kind of analysis and visualization, and required a summary of how the research answered or at least clarified the question or problem. It should be noted that this course did not teach research methods or disciplinary content knowledge, and the students needed to get content assistance from their own research group.

Student Assessment -- Results

Although in both course offerings several students indicated they had a rudimentary 'button pushing' understanding of the technologies probed in the pre-survey, none indicated that they felt able to command the tools to accomplish their own ideas and solutions. The survey, in fact, revealed low levels of exposure to most of the course content. They reported little experience with GIS at all, and what experience they had was limited to a handful of data types and rather turn-key operations. Both offerings of the course required the instructors to cover fundamental lessons before moving on to a higher-order agenda. These lessons included an introduction to databases and data formats, basic use of GIS and GPS tools, rudimentary visualization and analysis techniques, and metadata and presentation skills. The instructors decided against using some specific technologies because, for example, students had no experience working in Unix/Linux systems or using low-level programming languages.

Students indicated a high level of interest in all the topics covered in the class, including an appreciation for data information literacy skills. In the standard end-of-course evaluations (to which all students [N=12] responded), the course received an overall rating of 4.8 out of 5.0, and several students remarked that after taking the course they finally understood what they were doing and now

could contribute new procedures for analyzing data to their research groups. Five of twelve enrolled students completed the 2008 post-course survey, with the results summarized in Table 1.

The high level of interest in basic topics such as data formats and an introduction to databases indicate the relative lack of preparation in the core technology skills necessary to work in an e-research environment. Although all but one topic received a rating of at least 4.0 (Very Important), the lower rated topics did include the more traditional library-focused concepts, such as ontologies and data preservation.

Topic	Importance to Course	Importance to Research
Databases	4.8	5.0
Data Formats	5.0	4.8
Data Gateways/Portals	4.6	4.6
Introduction to GIS	4.8	4.8
GIS Analysis	5.0	5.0
GIS Data Conversion	5.0	5.0
Workflow Management	4.6	4.6
Metadata	5.0	5.0
Statistics	4.6	4.4
GPS	4.6	4.2
Data Visualization	5.0	5.0
Ontologies	4.0	3.6
Data Preservation	4.2	4.2

Table 1: Results of the 2008 post-course survey, on a 5-point Likert scale, of the importance of different topics to the course and to the students' research. (N=5)

In addition to extracting information from course surveys, the instructors also carefully examined students' completed coursework to determine which concepts, skills, or ideas students still lacked, even after their introduction to and practice during the course. For example, the authors found that most students had ready access to the primary data used by their research groups and that these data often formed the basis for their semester project analysis. A focus of the course was on students' ability to identify and synthesize supplementary data, such topographic, political, or land-use data to

overlay on the data collected by the research group. Analysis of the student semester projects indicated that students indeed could find, identify, and incorporate external data sources into their analysis and/or visualization.

However, the analysis of the students' semester projects from both years does reveal recurring shortcomings. While students did apply external data appropriately to their work, frequently, these data were not properly cited. A stream flow gauge web service would be cited as "Environmental Protection Agency," for example. Although students correctly documented traditionally published literature, it seems that students did not consider data to be a valid, citable scholarly source or did not have a clear understanding of how to cite a dataset.

Students also struggled to fully comprehend the importance and complexity of data sharing. This is no source of pride, given the course was in many ways geared toward pushing this point explicitly. It is a thorny, almost byzantine concept and a very tough sell to students in the span of a single semester. The following issues appeared multiple times over the two separate semesters:

1. Preservation / Archiving-- The students' final task in 2008 was to submit their data to the GEON Portal (www.geongrid.org) for safe-keeping and redistribution. In 2010, GEON was merely a suggestion and students were encouraged to identify a repository in their domain to which they could submit their project data. Although many students attempted these submissions in good faith (despite some technical difficulties with GEON both years), several students shared the sentiments of one in particular, who argued that a department-run website that "everybody in the [domain] community knows about" was a better ultimate destination for their data than any more formal data repository.

2. Metadata-- Although the time allocated for metadata was limited, the instructors managed to include the concepts of schema, authoritative terminology, xml, indexing, and searchability. Each course offering had a metadata unit during which instructors introduced students to several proper examples of metadata and then completed a lab in which they wrote their own simple metadata documents. While some students did indeed write good accompanying metadata for their final project materials, most did not. One deficit seemed to arise from students creating metadata from the perspective of "how I did it," rather than striving to make the data more discoverable by the next scientist down the line. A sample abstract from one student's DC.description.abstract field, for example, details that "This KML file was created as part of semester project for a student of Purdue University's EAS591 class. Using an old KML file from a previous lab session, the student modified the code to call a PHP file, which in turn calls a MySQL table. The contents of the MySQL are then put into Goggle Earth place marks by the KML file. The place marks display images, video, and text of a geologic nature."

3. The technologies and workflows of data sharing-- Students (despite instructor warnings) expected to accomplish far more than they were able during a single semester. The most common error of this sort was linked to the students' expectation that, once analyzed, their data could be fairly easily visualized and shared online. The complexity of building data-based, interactive web applications was not apparent until it was too late.

Discussion

The authors sought to triangulate the needs related to data information literacy through interviews with research faculty and by analyzing the results of our own geoinformatics-themed data information literacy course. We found a substantial amount of overlap between the needs identified using both methods: databases, metadata, data sharing, preservation and curation of data, and formatting and documentation of data.

The assessments also uncovered differences that, if not unique to each population, were more clearly a focus for one group than the other. For example, the interviews with faculty members primarily focused on data they create themselves, while a not insignificant portion of the geoinformatics course involved locating data from external sources. An analysis of course work showed that students needed to learn "the basics" of much of information technology, even before broaching data issues. Additionally, in order to manipulate the data, students had to learn how to use analysis and visualization tools, workflow management tools, and develop a minimum computing background to take advantage of the available cyberinfrastructure. On the other hand, the production- and publication-focused faculty researchers described the need for data curation and management, such as good versioning, documentation, quality assurance, and the merging of data. In addition, the faculty surfaced the concept of data ethics: when to share data, who owns data, and how to appropriately acknowledge data. To that extent, these two investigations provide complementary information about perceived data information literacy needs.

We have argued that a real-world understanding of either faculty or student practices and needs alone are insufficient to develop the foundational objectives necessary for a data information literacy program. Instead both faculty and student perspectives must be understood and analyzed in tandem to inform a more complete understanding of what is needed in data information literacy. We now reintroduce another foundational component towards developing objectives for a DIL program, the perspective of the librarian. The organization, description, dissemination, curation, and preservation of information resources, which increasingly includes research data, are the hallmark of librarians. Although DIL must be grounded in real-world needs as expressed by students and faculty, the librarian brings the broader perspective and a connection to the larger "information ecology" that exists beyond the single research project or classroom. This connection ensures that best practices inform current practices.

Comparison of Data Information Literacy with ACRL IL Standards

To help articulate and ground our core DIL objectives, the authors find it useful to examine these topics through the prism of our current literacy framework, the ACRL information literacy competency standards, which guide many library instruction initiatives.[xxx] To that end, the next section first lists the ACRL standards, then briefly examines each standard in turn for its relevance to these DIL objectives.

One readily identifiable gap in applying the ACRL information literacy standards towards developing a data information literacy program is the difference in focus. The ACRL standards focus on educating the information consumer, people seeking information to satisfy an information need. Although faculty and students do consume research data, our analysis of faculty and students indicates a strong need to address their roles as data producers as well. Therefore, the underlying objectives for any data information literacy program need to accommodate both the data producer's viewpoint as well as that of the data consumer.

The ACRL standards include:

1. Determine nature and extent of information need
2. Access needed information efficiently and effectively
3. Evaluate information and its sources critically and incorporates selected information into his or her knowledge base and value system.
4. Use information effectively to accomplish a specific purpose.
5. Understand many of the economic, legal, and social issues surrounding the use of information and accesses and uses information ethically and legally.[xxxi]

Standard One: Identifying nature and extent of information need

When gathering information, one often skips over the research question formulation stage that is the foundation of the information search process.[xxxii] However, without understanding the question deeply, one cannot arrive at a correct answer. The instructors addressed this concept in the semester project for the Geoinformatics class, for example, as the overall assignment asked students to first identify their research question and from there determine what data they needed in order to address that question. In the case of geospatial data, students needed to determine whether to use raster or vector data, because each type has its own strengths and weaknesses for analysis and presentation. Thus, the authors' curricular topic of databases and data formats fit best into this competency standard, as they are fundamental to understanding the nature of the information needed. In fact, Standard One already explicitly addresses data, stating that a student "realizes that information may need to be constructed with raw data from primary sources."

From the data producer's standpoint, identifying the nature and extent of the potential needs and uses of the data being generated provides the foundation for effectively sharing, re-using, curating and preserving of data. The cultural practices and norms of the producer's discipline, including an awareness of any existing community resources, standards, or tools, inform these data functions.

Standard Two: Access needed information efficiently and effectively

Students need to consult common disciplinary and general data repositories as well as understand the formats and services through which data can be accessed in order to access information efficiently and effectively. In the geoinformatics course, students investigated several data sources and were required

to use external data extensively to supplement their own data. In addition to finding data relevant to their research question, the variety and complexity of data formats made the process of locating supplementary data challenging for the students. Several students needed assistance converting data from one format to another and understanding how to merge datasets with different resolutions or timescales.

Standard Two addresses these issues, as an information literate student "extracts, records, and manages the information and its sources," including using "various technologies to manage information selected and organized." Not only will DIL students need to know where data exist, but they also must harvest, convert, possibly merge, and ultimately feed it into analysis or visualization tools that may or may not require still other formats. Although a direct graft of classic information literacy competency standards to our DIL would focus on the process of bringing data *into* one's research, as the faculty interviews revealed these concepts are similar for publishing data to the world. Thus, DIL concepts related to this competency standard include data repositories, data conversion, data organization, sharing data and interoperability.

Standard Three: Evaluate information critically

When evaluating data, information literate students understand and critically evaluate the source. Students must determine whether the research group that provided the data is reputable and/or if the data repository or its members provide a level of quality control for its content. Users also need to evaluate the data for relevancy and compatibility with their own research. In addition, and as part of the quality assurance component of data evaluation, students need to evaluate associated metadata. Among other attributes, metadata specifies the details of the experiment or data product, including the conditions under which the data were collected or created; the apparatus or procedures used to generate the data; distribution information and access rights; as well as spatial and temporal resolution, units, and parent sources. It is therefore a vital tool in the evaluation of the quality and authority of the resource. While the ACRL standards would approach this from a data user perspective, the faculty interviewed made it clear that data producers need to provide quality assurance for data and metadata as well.

Standard Four: Use information to accomplish a specific purpose

In this standard, students are assumed to be carrying out a research project and will need to "communicate the product or performance effectively to others." As such, students should use a format and employ appropriate information technologies that best support the purpose of the work. Here, in the expansive verb "communicate" and phrase "appropriate information technologies," one can find the concepts of data sharing, re-use and curation as well as connections to analysis and visualization tools.

In addition, this standard includes the application of information towards the planning and creation of a product, revising the development process as appropriate along the way. These components parallel the statements made by faculty on the importance of documenting the processes used to develop research data (the "product" in this case). Researchers also identified the careful management and organization of data as essential in enabling its eventual transfer "from their original locations and formats to a new context" (as stated in Standard Four) for internal use by others in the project, or for re-use by others.

Standard Five: Understand economic, legal, and social issues and use information ethically

Data ethics are certainly an important component of a well-rounded DIL program, especially since intellectual property issues around data are much more fluid than, for example, traditional textual works. Students need to not only determine when and how to share data, which varies among disciplines, but also document their own sources of data. We found students struggled with the latter in the geoinformatics course, as exhibited primarily by a failure to acknowledge those parties responsible for the data they consumed and re-used. The ethical issues surrounding students as data producers and publishers, a concern raised by research faculty, appears to be entirely absent from the ACRL standards and would be a largely novel component of a DIL curriculum.

Core Competencies for Data Information Literacy

Based on information gleaned from the faculty interviews, geoinformatics course, and ACRL Information Literacy Competency Standards, the authors propose the following educational objectives for a data information literacy program, understanding that disciplinary variations of these outcomes would incorporate technologies or techniques specific to that discipline. The proposed core competencies, organized by major theme, are listed below.

- **Introduction to Databases and Data Formats**
  Understands the concept of relational databases, how to query those databases, and becomes familiar with standard data formats and types for their discipline. Understands which formats and data types are appropriate for different research questions.
- **Discovery and Acquisition of Data**
  Locates and utilizes disciplinary data repositories. Not only identifies appropriate data sources, but also imports data and converts it when necessary, so it can be used by downstream processing tools
- **Data Management and Organization**
  Understands the lifecycle 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.
- **Data Conversion and Interoperability**
  Becomes 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.

- **Quality Assurance**
  Recognizes and resolves any apparent artifacts, incompletion, or corruption of data sets. Utilizes metadata to facilitate understanding of potential problems with data sets.
- **Metadata**
  Understands the rationale for metadata and proficiently annotates and describes data so it can be understood and used by themselves, others in their workgroup, and external users. 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.
- **Data Curation and Re-use**
  Recognizes that data may have value beyond their original purpose, to validate research or for use by 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 lifecycle. Articulates the planning and actions needed to enable data curation.
- **Cultures of Practice**
  Recognizes the practices, values, and norms of his/her chosen field, discipline or sub-discipline as they relate to managing, sharing, curating and preserving data. Recognizes relevant data standards of his/her field (metadata, quality, formatting, etc.) and understands how these standards are applied.
- **Data Preservation**
  Recognizes the benefits and costs of data preservation. Understands the technology, resource and organizational components of preserving data. Utilizes best practices in preservation appropriate to the value and reproducibility of data.
- **Data Analysis**
  Becomes familiar with the basic analysis tools of their discipline. Uses appropriate workflow management tools to automate repetitive analysis of data.
- **Data Visualization**
  Proficiently uses basic visualization tools of discipline. Avoids misleading or ambiguous representations when presenting data. Understands the advantages of different types of visualization, for example, maps, graphs, animations, or videos, when displaying data.
- **Ethics, including citation of data**
  Develops an understanding of intellectual property, privacy and confidentiality issues, and the ethos of their discipline when it comes to sharing data. Appropriately acknowledges data from external sources.

By way of corroboration, the authors compared the DIL core objectives with the course syllabus from the Science Data Literacy curriculum of Qin and D'Ignazio[xxxiii] and found a large degree of overlap between the two formulations. The chief difference appears to be the depth of treatment of different topics. While the SDL course concentrates on metadata, for example, our approach focuses as much on the consumption of data (tools) as it does on documenting and annotating data. Even with a standing admission that the geoinformatics course perhaps had too little coverage of metadata, we found that students and faculty both need just as much help with data manipulation as they do with enhancing the documentation of their data. Naturally, instructors must determine the right balance between using tools and creating interoperable infrastructure in their own offerings of this type of course.

We have alluded already to the notion that a fully functional, richly-stocked DIL program may not be entirely under the domain of librarians. However, if a librarian does have the skills required to teach database management and data analysis, for example, there is no reason why they should not teach those concepts. Indeed, learning those skills can help librarians remain integral to the educational mission of the university. However, the authors do recommend a collaborative venture between disciplinary faculty and librarians as the best practice for teaching data information literacy skills. DIL needs to be grounded in the culture of the discipline in which it is embedded, to be sure, but certainly imbued with the greater, communal perspective possessed by an inter- or extra-disciplinary librarian.

Conclusion

> ” …[T]hirty years ago, it was good laboratory practice [that] you had a bound paper manual, you took good notes, you took fifteen or twenty data points, maybe a hundred, and you had a nice little lab book. But we’ve scaled now to getting this mega amount of information and we haven’t scaled our laboratory management practices…It makes perfect sense to me that…you get this [data management skills] in people’s consciousness, make them aware it’s a problem early on in their careers as graduate students, before they go on and do all the other things and get too set in their ways… And…that takes a fair amount of education…and training.” (Civil Engineer)

The authors have uncovered a growing need among research faculty and students for data information literacy skills. As a result, the authors brought together data from different audiences to propose a suite of core data information literacy skills that future e-researchers need in order to fully actualize the promise of the evolving cyberinfrastructure.

Data information literacy represents an opportunity to expand information literacy from the library into the laboratory. In much the same way that libraries’ information literacy programs have gone beyond the “one-size fits all” approach, librarians will need to go beyond a “one shot/one-size fits all” approach to data management and curation literacy. The Data Curation Profiles project[xxxiv] indicates that different disciplines and sub-disciplines have different norms and practices for conducting their research and working with their data. These differences are manifest in the myriad ways they manage (or don't manage), share (or don't share), curate and preserve (etc.) their research data. While we have provided a general summary of common themes from these interviews, we understand that any DIL program focused on a specific discipline needs to be able to identify, incorporate, and address these specific differences into the curricula. Models are needed to help ascertain the educational needs of sub-disciplines with regard to their data and then design DIL programs that will address these needs. To that end, these results serve to start a conversation, and propose general concepts, rather than provide a final, detailed curriculum.

Upon examination of the ACRL standards for information literacy, it becomes clear that data information literacy does fall within the scope of previous library practice. The conceptual overlap between the ACRL standards and the DIL objectives indicates that these kinds of skills are very much in

the jurisdiction of librarianship. With some exceptions, the ACRL standards are written generally enough to accommodate DIL skills, and indeed the standards do have several specific outcomes related to data. Still, given the ballooning interest in data management for e-research, there may be a need to re-examine those standards and incorporate more data-related outcomes, especially from the perspective of the user as publisher and not just consumer of information. Additional scholarship can surely suffuse current standards with these developing perspectives and expectations.

Additional research should be done to map the skill sets librarians currently have to the data information literacy objectives, either as stated here or as they develop in practice. This will speed the development of not only a new DIL curriculum, but also push the community to continuously work to adapt the collective DIL practice to emerging trends in e-research.

Acknowledgements

The authors wish to acknowledge support from the Institute of Museum and Library Services, Grant #IMLS LG-06-07-0032-07 for the faculty assessment portion of this paper, and the work of Anupma Prakash, University of Alaska-Fairbanks, whose own geoinformatics course provided insight into the development of our own.

---

[i] Elisabeth Jones et al., *E-Science Talking Points For ARL Deans And Directors* (Association of Research Libraries, October 2008), http://www.arl.org/bm~doc/e-science-talking-points.pdf (accessed October 9, 2010).

[ii] National Science Foundation, "Grant Proposal Guide 2011," (2010), http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp (accessed October 10, 2010).

[iii] National Virtual Observatory, "What is the NVO?," *Welcome to the US National Observatory*, http://www.us-vo.org/what.cfm (accessed October 10, 2010); Jim Gray et al., "Online scientific data curation, publication, and archiving," Arxiv preprint cs/0208012, (Microsoft Research, Microsoft Corporation, 2002), http://research.microsoft.com/pubs/64568/tr-2002-74.pdf (accessed October 11, 2010).

[iv] The Monk Project, "Metadata Offer New Knowledge (MONK)," http://monkproject.org (accessed October 10, 2010).

[v] Daniel E. Atkins et al*., Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure* (Arlington, Va: National Science Foundation, 2003), http://www.nsf.gov/od/oci/reports/toc.jsp (accessed October 6, 2010).

[vi] National Science Board, *Long-lived Digital Data Collections Enabling Research and Education in the 21st century* (Washington, D.C.: National Science Foundation, 2005).

[vii] Commission on Cyberinfrastructure for the Humanities and Social Sciences, *Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences* (New York, NY: American Council of Learned Societies, 2006).

[viii] Committee on Science and Public Policy (U.S.), *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age* (Washington, D.C.: National Academies Press, 2009).

[ix] Tony Hey and Jessie Hey, "e-Science and its Implications for the Library Community," *Library Hi Tech*, 24, 4 (2006): 515 – 528; 517.

[x] Ibid., 517.

[xi] Anna Gold, "Cyberinfrastructure, Data, and Libraries, Part 2: Libraries and the Data Challenge: Roles and Actions for Libraries," *D-Lib Magazine* 13, 9/10 (2007), doi:10.1045/september2007-gold-pt2. [Note this journal does not have page numbers.]

[xii] Joint Task Force on Library Support for E-Science, *Agenda for Developing E-Science in Research Libraries* (Washington, DC: Association of Research Libraries, 2007), http://www.arl.org/bm~doc/ARL_EScience_final.pdf (accessed October 12, 2010).

[xiii] Tyler O. Walters, "Data Curation Program Development in U.S. Universities: The Georgia Institute of Technology Example," *The International Journal of Digital Curation* 4, 3 (December 2009): 83-92; 84, http://ijdc.net/index.php/ijdc/article/view/136/0 (accessed October 12, 2010).

[xiv] Joyce L. Ogburn. "The Imperative for Data Curation." *portal: Libraries and the Academy*, 10, 2, (April 2010): 241-246; 244.

[xv] Amy Friedlander. *To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering* (Washington, DC: Association of Research Libraries, 2006): 122.

[xvi] National Science Foundation, Cyberinfrastructure Council, *NSF's Cyberinfrastructure Vision of 21st Century Discovery* (Washington, DC: National Science Foundation, 2005), 38, http://www.nsf.gov/od/oci/CI-v40.pdf (accessed October 12, 2010).

[xvii] Tracy Gabridge. "The Last Mile: Liaison Roles in Curating Science and Engineering Research Data," *Research Library Issues: A Bimonthly Report from ARL, CNI, and SPARC* 265 (August 2009): 15–21; 17, http://www.arl.org/bm~doc/rli-265-gabridge.pdf (accessed October 12, 2010).

[xviii] Milo Schield, "Assessment Methods in Statistical Education: An International Perspective," *Assessing Statistical Literacy: Take CARE*, edited by Penelope Bidgood, Neville Hunt, and Flavia Jolliffe (Wiley: NY, 2010): 135.

[xix] Milo Schield. "Information Literacy, Statistical Literacy and Data Literacy," *IASSIST Quarterly* (Summer/Fall 2004): 7-11; (All quotations, p. 8.).

[xx] Elizabeth Stephenson and Patti Schifter Caravello,"Incorporating Data Literacy Into Undergraduate Information Literacy Programs in the Social Sciences," *Reference Services Review*, 35, 4 (2007): 525-540.

[xxi] Jian Qin and John D'Ignazio, "Lessons Learned from a Two-year Experience in Science Data Literacy Education," in *Proceedings of the 31st Annual IATUL Conference, June 20-24, 2010*. IATUL (West Lafayette, Indiana: IATUL, 2010): 2, http://docs.lib.purdue.edu/iatul2010/conf/day2/5 (accessed October 12, 2010).

[xxii] Anne Graham, Amy Stout and Katherine McNeill, "Managing Research Data 101," MIT Libraries (2010), http://libraries.mit.edu/guides/subjects/data-management/Managing_Research_Data_101_IAP_2010.pdf (accessed October 12, 2010); LibGuides page available at http://libraries.mit.edu/guides/subjects/data-management/ (accessed October 12, 2010).

[xxiii] Carol Hunter et al., "A Case Study in the Evolution of Digital Services for Science and Engineering Libraries," *Journal of Library Administration* 50, 4, (2010): 335. doi:10.1080/01930821003667005

[xxiv] Jian Qin and John D'Ignazio, "The Central Role of Metadata in a Science Data Literacy Course," *Journal of Library Metadata* (in press). See also "The Science Data Literacy Project," http://sdl.syr.edu (accessed October 15, 2010).

[xxv] Science Data Literacy Project, "Syllabus," http://sdl.syr.edu/?page_id=23 (accessed October 15, 2010).

[xxvi] C.C. Miller and Michael Fosmire, "Creating a Culture of Data Integration and Interoperability: Librarians Collaborate on a Geoinformatics Course," in *Proceedings of the 29th Annual IATUL Conference, April 21-24, 2008*. IATUL (Auckland, New Zealand: IATUL, 2008), http://www.iatul.org/doclibrary/public/Conf_Proceedings/2008/MFosmire080320.doc (accessed October 12, 2010).

[xxvii] C.C. Miller, "Geoinformatics Course Site," (April, 2010), http://www.lib.purdue.edu/gis/geoinformatics (accessed October 15, 2010).

[xxviii] Michael Witt et al., "Constructing Data Curation Profiles," *International Journal of Digital Curation* 4, 3 (2009): 93-103, http://www.ijdc.net/index.php/ijdc/article/viewFile/137/165 (accessed October 15, 2010).

[xxix] Deborah J. Grimes and Carl H. Boening, "Worries about the Web: A Look at Student Use of Web Resources," *College and Research Libraries* 62, 1 (2001): 11-23.

[xxx] American Library Association, Association of College and Research Libraries, *Information Literacy Competency Standards for Higher Education.* (Chicago, IL: ACRL, ALA 2000), http://www.ala.org/ala/mgrps/divs/acrl/standards/standards.pdf (accessed October 9, 2010).

[xxxi] Ibid., 2-3.

[xxxii] Carol Collier Kuhlthau, *Seeking Meaning: a Process Approach to Library and Information Services.* 2nd ed. (Westport, Conn.: Libraries Unlimited, 2004).

[xxxiii] Ibid., Jian Qin and John D'Ignazio. *Journal of Library Metadata*.

[xxxiv] Ibid., Witt, et.al. 2009; Melissa Cragin et al., "Data Sharing, Small Science, and Institutional Repositories," *Philosophical Transactions of the Royal Society A.* 368 (2010): 4023-4038.