An Exploratory Analysis of Social Science Graduate Education in Data Management and Data Sharing

Ashley Doonan
ICPSR, University of Michigan

Dharma Akmon
ICPSR, University of Michigan

Evan Cosby
ICPSR, University of Michigan

Abstract

Effective data management and data sharing are crucial components of the research lifecycle, yet evidence suggests that many social science graduate programs are not providing training in these areas. The current exploratory study assesses how U.S. masters and doctoral programs in the social sciences include formal, nonformal, and informal training in data management and sharing. We conducted a survey of 150 graduate programs across six social science disciplines, and used a mix of closed and open-ended questions focused on the extent to which programs provide such training and exposure. Results from our survey suggested a deficit of formal training in both data management and data sharing, limited nonformal training, and cursory informal exposure to these topics. Utilizing the results of our survey, we conducted a syllabus analysis to further explore the formal and nonformal content of graduate programs beyond self-report. Our syllabus analysis drew from an expanded seven social science disciplines for a total of 140 programs. The syllabus analysis supported our prior findings that formal and nonformal inclusion of data management and data sharing training is not common practice. Overall, in both the survey and syllabi study we found a lack of both formal and nonformal training on data management and data sharing. Our findings have implications for data repository staff and data service professionals as they consider their methods for encouraging data sharing and prepare for the needs of data depositors. These results can also inform the development and structuring of graduate education in the social sciences, so that researchers are trained early in data management and sharing skills and are able to benefit from making their data available as early in their careers as possible.

Correspondence should be addressed to Ashley Doonan, 330 Packard Street, Ann Arbor. MI 48104. Email: ebashley@umich.edu

Portions of this paper were presented at the annual conferences of the International Association for Social Science Information Services and Technology (IASSIST), in Lawrence, KS in 2017, and in Montreal, Canada in 2018.

NOTE: This working paper is circulated for discussion and comment purposes. It has not been peer-reviewed but may be under review. This working paper was submitted to the *International Journal of Digital Curation* for peer review.

Introduction

Research training is a crucial part of social science graduate programs which aim to prepare students to be well-rounded scholars and professionals. Although the majority of these programs offer curricula in research methods and provide experience in data collection, it is less clear how many students leave graduate school prepared to effectively manage and share their data. Effective data management make data archival and reuse possible, and many sponsors now include data dissemination or archiving as a requirement for funding. In some cases, researchers are legally required to maintain their research data well after funding for their project is complete (Zacharias, 2010). Research suggests that social science researchers do not feel satisfied with their training in data management, with much of their training occurring on the job rather than through formal education (Jahnke & Asher, 2012). Whether emerging scientists are being effectively taught the specific skills necessary to manage and share data remains unclear.

Education can be divided into three major modes according to Coombs and Ahmed's framework: formal, nonformal, and informal (1974). They define each of the three modes as follows: formal education is the structured and institutionalized method of learning through schooling and coursework; nonformal education is any organized educational activity occurring outside of the formal school system; and informal education is the unorganized process of learning from experience and exposure in the environment, such as through reading or peer example. This could include training programs, occupational skill training, community programs, or activities such as educational clubs. These three types of learning can occur simultaneously in nuanced ways, both as distinct modes of education (the learning process) and as characteristics of learning (the structure of the education) (La Belle, 1982).

Research from Jurić (as cited in Dikovic & Plavsic, 2015) suggests that most learning is informal, and Coombs and Ahmed posit that both formal and nonformal education systems exist to supplement and expand upon informal learning (1974). Certain types of knowledge and skills that are not readily or quickly acquired through typical informal exposure, such as work-related skills, depend on formal or nonformal education (Coombs & Ahmed, 1974; La Belle, 1982). This is especially true for learning effective data management across the research lifecycle, which can be very challenging given the complex variety of skills, resources, and knowledge required (Hou et al., 2017). Large amounts of data and varying formats and file types make data management all the more challenging (Tenopir et al., 2016). With trends toward larger-scale data collection, and the hurried pace of technology, data management is set to become even more important to know, but more difficult to master.

Graduate programs place significant emphasis on formal coursework, nonformal learning through theses and research, and informal professional experiences such as conferences and presentations as a way for to develop skills for conducting research. However, it is unclear if this same emphasis is placed on the research skills needed for sharing and managing data, such as mitigating disclosure risk, maintaining dataset integrity, and describing data so others can understand them. Evidence shows that many researchers do not consider long-term preservation when conducting research (Jahnke & Asher, 2012) nor do they properly prepare their data for sharing (Savage & Vickers, 2009), despite the critical role effective data management plays in the research lifecycle

and funder requirements to do so. An analysis of funded NSF Data Management Plans found that data sharing plans showed little understanding of how to share data in a way that meets public access requirements (Bishoff & Johnston, 2015). This implies a lack of knowledge or sufficient training in data management and archival skills. Informal learning, especially in the workplace, requires a large investment of effort and must be initiated by the individual, meaning deterrents like expense and lack of time play a major role (Lohman, 2005). This results in learning only occurring as knowledge and skills are needed. Without a strong foundation of formal training experiences, many researchers are unprepared to manage data for later archiving and reuse, and often at a time when they are short on time and/or funding.

In their survey of social scientists' data curation practices, Jahnke and Asher found that researchers received scant formal training in data management practices (2012). Respondents reported that what little training they did receive was a cursory part of research methods courses. Some participants reported seeking out informal or nonformal supplemental training such as consulting with experts or books. This survey, however, consisted primarily of professors and did not elucidate current practices of graduate programs in providing data management training. A more recent survey of research educators suggested the majority of instructors felt they were not teaching data management topics sufficiently, and nearly a third of surveyed educators reported they were only teaching data management outside the classroom environment (Tenopir et al., 2016).

Data and information professionals are filling some of the training gaps through structured supplemental trainings, such as workshops, online training, and even some courses (Carlson & Stowell Bracke, 2015). Student evaluations of data curation graduate courses suggest that students feel they benefit greatly from structured learning of data management skills (Kelly et al., 2013). However, these primarily nonformal and informal supplements cannot completely make up for a lack of formal training. One challenge of relying on supplemental, nonformal trainings provided by data and information professionals is that they are often not directly relevant to the student's discipline or unique needs (Carlson & Stowell Bracke, 2015). Many graduate students view courses outside of their discipline or program requirements as a distraction (Carlson, Johnston, Westra, & Nichols, 2013), although graduate school represents the most opportune time for learning a discipline's norms and practices. Less time-intensive options, such as workshops, reach more graduate students but cannot provide the same depth of training as a course (Carlson et al., 2013), highlighting the need for some level of discipline-specific, formal training.

Our study assesses how masters and doctoral level social science programs include formal, nonformal, and informal learning in data sharing and data management. For this research, learning was defined in the following way: a) formal training was defined as any learning and instruction that occurs within the classroom setting, b) nonformal training was any learning with a defined goal that was a requirement or part of earning the graduate degree, but did not occur within a classroom, and c) informal training was any experience based learning that had no clearly defined goals but might occur as a result of being in the graduate program. We aimed to discover what content is included in social science research training and especially to what extent students receive training in data management. This study also explored whether students are exposed to data sharing or secondary analyses in their training. This analysis specifically explores current graduate program practices to gain insight into what knowledge early career researchers are equipped with as they leave graduate school.

Methods and Results

We conducted this study in two stages. First, we carried out a survey of a select set of social science graduate programs to assess the inclusion of data management and sharing-related content, as well as the experiential development of these skills as a result of participating in the program. For the survey, we wanted to determine if content was included in programs formally, nonformally, or informally, or if it was absent altogether. The characteristics we used to conceptualize graduate learning and education for the present study are detailed in Table 1. Following the completion of the survey, we conducted a syllabus analysis to further explore explicit inclusion of formal data management and data sharing training. We aimed to determine more clearly through the syllabus analysis the extent of formal versus nonformal content inclusion.

Table 1. Characteristics used to define the three modes of learning.

Formal learning	Nonformal learning	Informal learning
Structured	Semi-structured	Unstructured
Planned	Planned	Unplanned
Compulsory	Compulsory or Voluntary	Voluntary
Intentional	Deliberate	Incidental
Institutionalized	Organized, out of school	Experiential
Hierarchical	Systematic	Unsystematic
Officially sanctioned	Can be sanctioned	Exposure-based

To determine what specific disciplinary fields to include, we utilized a broad definition of social science and examined the ethical codes established by American governing bodies for different social science disciplines. We limited our study to disciplines whose governing body included data sharing or open data access requirements in their published code of ethics. This resulted in the survey sampling from programs from six disciplines: anthropology, history, geography, psychology, sociology, and political science. Between the completion of our survey and the development of our sampling frame for the syllabus analysis, the governing body for economics, the American Economic Association (AEA), released a draft of their code of conduct (American Economic Association, 2018). Because this code of conduct mandated research transparency, we felt it appropriate to include economics in addition to the other six fields.

The University of Michigan Health Sciences and Behavioral Sciences Institutional Review Board deemed this study exempt from ongoing IRB review. No program identifying questions were asked, and this paper reports the findings in the aggregate. We obtained informed consent for survey participation prior to presenting participants the survey questions.

Survey of Graduate Programs

We conducted our survey of social science graduate programs utilizing Qualtrics Survey Software. We developed a sampling frame of randomly selected graduate programs from our six selected social science fields using the graduate school directory gradschools.com. We restricted our sampling to in-person, research-based programs in the United States, but included masters and doctoral level programs at both public and private institutions across all Carnegie classification levels. We gathered a simple random sample of 25 programs matching these criteria from each of the six disciplines. After identifying programs, we sent the survey to program directors, administrators, and other representatives designated as the contact person for the program on their university website. Ultimately, we sent our survey to 150 graduate programs and received responses from 27 (18%). The final dataset included 24 usable responses.

The survey consisted of questions about program demographics, the inclusion of specific coursework, student research participation, and student learning and post-graduation outcomes. To better understand data management and sharing education at respondents' institutions, the survey included questions about program courses, specifically exploring the inclusion of research methods, data sharing, data management, and research ethics. Additional items asked respondents to provide information about if and how students are given data sharing and data management training or information. The section of items on student research participation asked about theses and dissertations, involvement in data collection, statistical package experience, publications, presentations, assistantships, and data repository use. Table 2 organizes the survey question topics into categories according to learning type. The final sections of the survey asked respondents to assess graduates' skills in research and data management and sharing upon graduation.

Table 2. Survey question topics sorted by learning type.

Formal learning	What courses are offered/required Content included in classes
Nonformal learning	Required research projects Mandatory statistical packages Theses/dissertations Data collection Supplemental class material
Informal learning	Student publishing Research assistantships Mentoring/Advising Data repository use Poster presentations or oral presentations

Survey Results

Response rates varied between programs, and only political science had a response rate greater than 25% (28%, 7) (Figure 1A). Our respondent pool was primarily comprised of program directors (41.7%, 10) and program coordinators (33.3%, 8). Other respondent positions included department chairs, program faculty, and academic

coordinators (Figure 1B). The 24 programs averaged 54 students (M = 54.33, SD= 75.61) for typical enrollment, and the number of full-time faculty in these programs ranged from two to 35, with a median of seven faculty members. The majority of programs offered both masters and doctorate degrees (62.5%, 15); seven programs (29.2%) offered only terminal masters, and two programs offered only the doctoral option (8.3%).

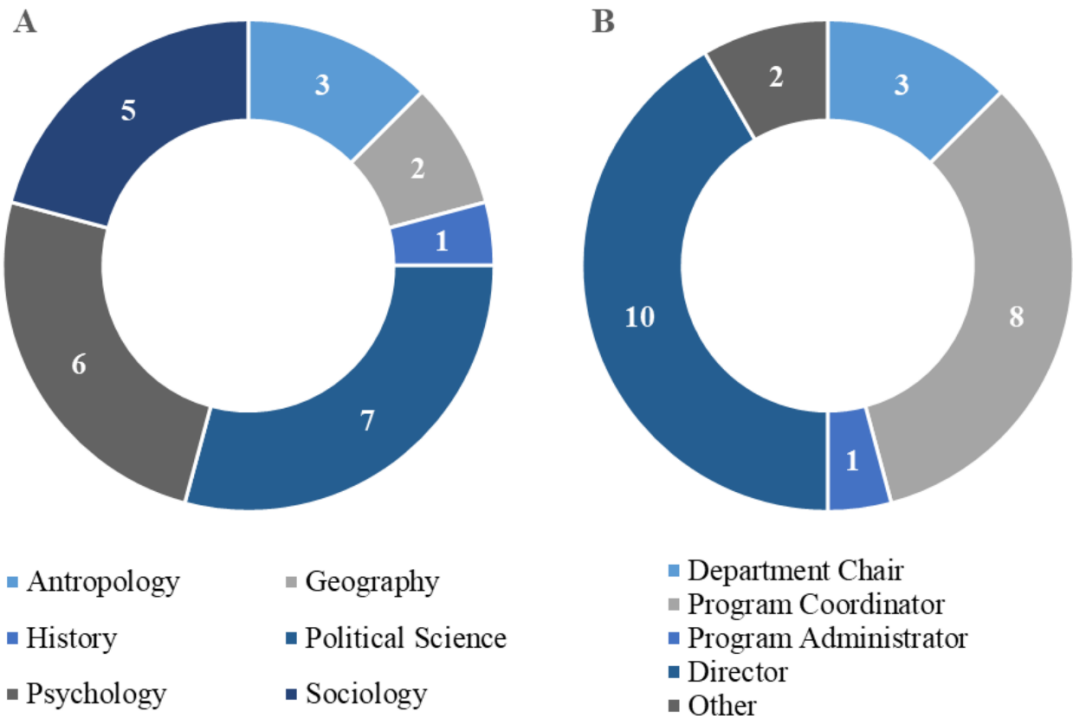


Figure 1. (A) Proportion of programs representing each of the selected social science fields in the survey responses. (B) Distribution of categorized respondent role in the graduate programs surveyed.

Program assessment of student skills

To determine how well programs felt they were preparing their students for research-based careers, we included questions in the survey assessing respondents' confidence level in their graduates. The majority of programs (16, 66.7%) in the survey expressed confidence in the adequacy of their graduates' research skills. However, when asked if they believed their students leave the program with a good understanding of data sharing, half of respondents reported being unsure (12), while 29.2% (7) reported that they do not think their graduates leave with a solid grasp on this topic. When asked whether they felt their students were graduating with good data management skills, nearly half of respondents reported either that they did not feel confident or that they were uncertain of their student' skills in this area (four [16.7%] did not feel confident; seven [29.1%] were uncertain). Four of these programs also reported mixed results among their graduates (e.g. "some do, some don't"), which may be an indication of some students pursuing supplemental material and learning or some difficulty or insufficiency of existing training.

Formal learning of data management and sharing

While programs reported fairly high levels of confidence in their graduates' research skills after graduation, when examining the formal content inclusion of data management and data sharing, results suggested little formal training. Although some programs reported that data sharing and data management content was covered formally through required courses, overall results suggested that if students are receiving this training, it is not in a structured fashion.

Nearly all programs that responded to the survey indicated that a research methods course was required as a condition for completing the degree (23 programs, 95.8%), showing consistent emphasis on teaching research skills. However, data sharing training content was sparse or overlooked. We also asked programs if their research methods course included either data sharing or data repository information. Just 37.5% (9) of programs reported covering information on both data sharing and data repositories in their research methods courses, and an equal number (9, 37.5%) reported not covering either of these topics at all. None of the respondents reported that their programs covered only data sharing (and excluded information about data repositories) in their research methods course, while only 12.5% (3) said they only included information about data repositories and not data sharing. A follow-up question asked whether graduate students were taught about specific data repositories they might use, and less than half of programs (11, 45.8%) reported providing such information in any coursework. Another 10 programs (41.7%) specified that information about data repositories was not covered at all. Not covering this information suggests that graduate programs place primary emphasis on learning how to collect new data, and the potential for reusing data is not formally discussed.

Our results suggest that the majority of formal data management training students receive comes from research methods courses. Of the 23 programs which included a research methods course, 75% (18) reported that this course included data management. It is encouraging that the majority of programs are including some degree of formal instruction on data management, however the depth of coverage possible in such a broad course is unclear. Despite the importance and breadth of information a researcher should learn about data management to be effective and fully prepared for responsible research, only one of the programs (4.2%) surveyed reported offering a data management course. Two programs (8.3%) reported that they planned to offer a data management course at some point in the future. Among the programs that did not offer a course specifically in data management, four (16.7%) reported that they did not provide information or training about data management in any other way. In other words, the only avenue for students in those four programs to learn data management would be to seek training on their own. Another eight (33.3%) responded that they did not know if information about how to manage data was shared with students outside of class, suggesting that this important training could be completely absent from their programs unless addressed in research methods courses. Social science programs may be relying too heavily on the content of their research methods courses to cover data management and data sharing.

Given that the ethical codes for the social science programs included in the survey mandated data sharing, graduate education for these fields would ideally contain formal ethics training with information about the importance of sharing data. However, only 20.8% (5) programs reported that they required a discipline-specific ethics course. Looking at these five programs, all indicated that research ethics is included within this course. Four (16.7%) programs reported that these courses included material from or reference to the ethical code from the governing body within the course. Verbatim

responses from three participants (12.5%) indicated that ethics was incorporated as part of another course or courses.

Nonformal learning of data management and sharing

We observed minimal inclusion of formal learning of either data management or sharing in our survey results. While formal learning is the most effective training method, we anticipated that graduate programs would include some semi-structured, nonformal learning experiences—such as research projects and theses/dissertations which require collecting data, to help students gain data management and data sharing skills. Our survey assessed nonformal learning opportunities in graduate programs by asking about semi-structured experiences that occur as part of program participation. We included significant projects completed outside of class, as well as any supplemental exposure to materials such as readings assigned or materials provided.

Nearly three quarters of programs surveyed (17, 70.8%) reported that students are required to complete a research project that involves collecting data, and this included theses and dissertations. For programs where data collection is not expected, interacting with data repositories could help students gain experience in reusing or manipulating existing data. Data repositories facilitate the ability to carry out analyses without collecting new data, so we anticipated that some programs would leverage such repositories. However, for nonformal training, only one program (4.2%) reported providing their graduate students with information about data repositories to review on their own outside of coursework. Similarly, looking only at programs who reported data management and data sharing not being covered in classes, we found that 16.7% (4) of respondents provided data management and data sharing information outside of class. These results suggest that students are not being encouraged to share the data they have collected after their research projects are completed.

Students' training in statistical packages was also explored in the survey as a potential indicator of areas where student may be receiving semi-structured learning about how to structure, organize, and maintain data. Most programs (21, 87.5%) reported that students use a statistical package to complete research as part of the program. Slightly less than half of programs (11, 45.8%) allowed their students to choose what statistical package they use, and only 16.7% (4) required the use of a specific package. This suggests less directed learning and training in how to use the statistical packages, especially for data management. In contrast, 29.2% (7) of programs reported that the statistical package requirement is dependent on the instructor. Lack of directed training may imply a reliance on undergraduate training or other independent experiences, such as informal learning or supplemental nonformal training sought out by the student. Such nonformal training experiences are unlikely to leave students with adequate skills for future data sharing and data management.

Informal learning of data management and sharing

Our survey findings indicate that both formal and nonformal data management and sharing instruction are lacking in social science graduate training. Our survey also explored whether students are learning these skills through informal means – those experiences where students acquire skills and knowledge as a function of being in graduate school, and not from any direct, intentional interventions by the program. We considered informal experiences to be unstructured and unplanned, experiential learning – skills which are gained despite that they are not explicitly taught. Several programs indicated that graduate students learned data sharing and data management skills through research assistantships, internships, or mentorship from their advisors. Among

programs that reported their students learned data management and data sharing skills from information provided outside of classes, five different open-ended responses indicated assistantships and advising as means for gaining this knowledge without explicit program intervention.

We examined students' informal data management learning experiences gained through practice and exposure. One key opportunity is through creating scholarly work such as publications and presentations, where students may be exposed to data management and sharing. Programs were asked to rank from 0 to 100 what percentage of their students typically engaged in publication experiences while enrolled. Many programs (20, 83.3% of respondents) reported that at least one percent of their students publish papers as first authors while enrolled. The average rate of independent publication reported was 34% (M = 34.00, SD = 28.07). Additionally, 20 programs (83.3%) indicated that at least some students publish as a coauthor with a professor. These 20 programs reported an average of 46.6% of students publishing with a faculty coauthor (M = 41.60, SD = 32.15). Looking at additional experiences of giving presentations, both posters and talks, programs reported an average of 66% of students giving posters (M = 66.33, SD = 26.92) and an average of 63% of students giving talks at conferences (M = 63.18, SD = 30.54). Table 3 outlines the average percentage of students who participate in these scholarly experiences in graduate school which might lead to informal learning of data management or data sharing skills.

Table 3. Informal Exposure to Sharing and Management by Scholarly Experience Type

Scholarly Experience	Minimum	Maximum	Mean	SD
Percentage of students that publish independently	0%	80%	34.0%	28.07
Percentage of students that publish as a coauthor with a professor	2%	95%	41.6%	32.15
Percentage of students that give talks at conferences	5%	100%	63.2%	30.54
Percentage of students that give poster presentations	10%	100%	66.3%	26.92

In terms of informal exposure to data repositories, where students would be exposed to the practice of sharing data and ease of secondary data use, student experience appears to be more limited. Half of programs (12) reported that their students use data repositories for any reason. Programs were then asked to select all the reasons which best describe why their students use data repositories, and these responses are outlined in Figure 2. Ten programs (41.7%) reported that their students use repositories to acquire supplemental data, and eight programs (33.3%) reported students use repositories to obtain bibliographic references. Only two programs reported that students use repositories to share their own data, and no programs reported that students share the data of an advisor or other faculty.

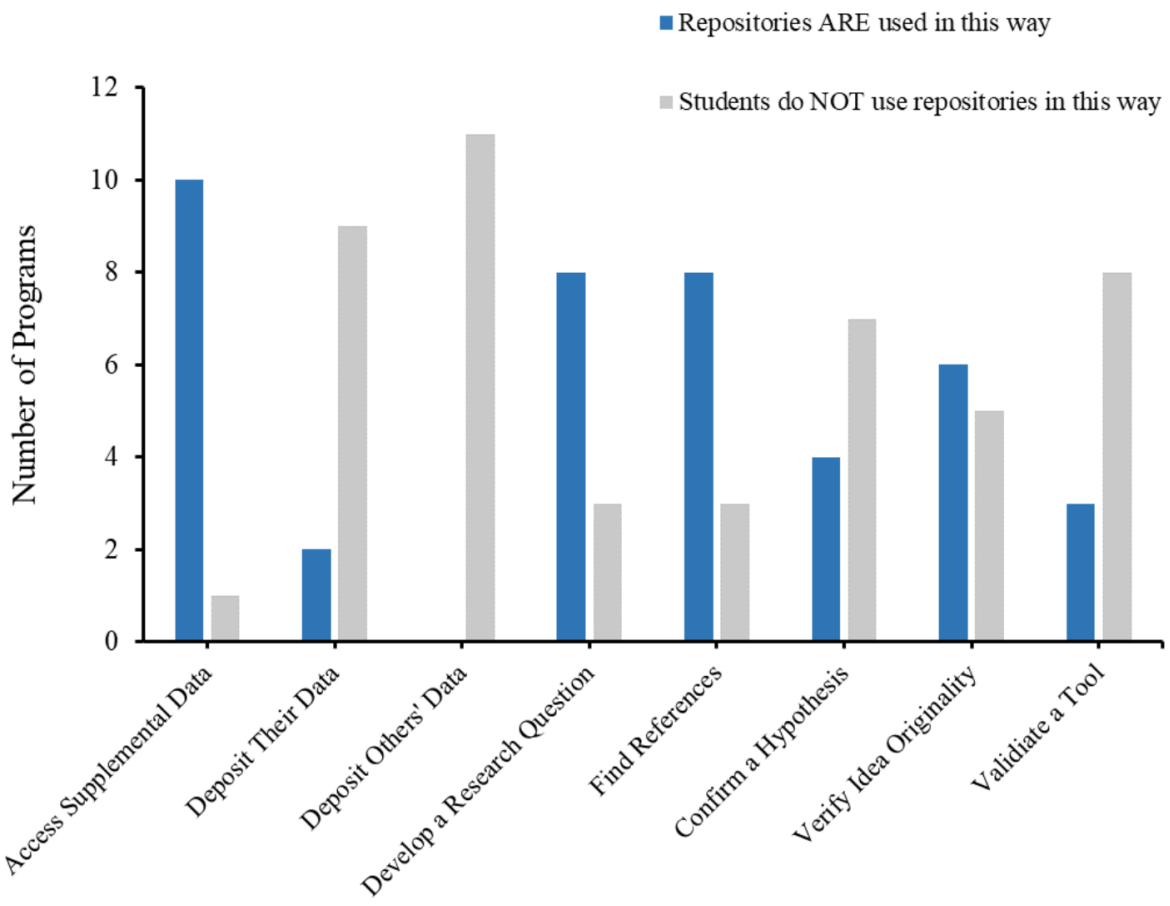


Figure 2. Distribution of how many programs did versus did not indicate that their students are using data repositories for a given reason.

Syllabi Analysis

To further explore formal and nonformal content included in graduate training, we followed the program survey with an analysis of syllabi from similar programs. Because our survey suggested that graduate programs are primarily utilizing the content of their research methods courses to teach their students about data management, we sought to explore how prominently this material might appear in the syllabi for this coursework, if at all. If data management and data sharing appears within a course syllabus, this would suggest greater instruction time devoted to the material. We constructed our sampling frame of graduate programs using the same method and inclusion criteria we employed for the survey. We randomly sampled programs within each of the disciplines, selecting 20 from each of the seven fields, for a total of 140 programs. These programs were diverse in size and format, and sampling resulted in representation from across all U.S states. For each program, we identified one course from their listed degree requirements, and we attempted to obtain a syllabus for these courses. Selected courses had to be relevant to learning how to conduct research. These included courses like Research Methods or Methodology, Statistics or Data Analysis, Research Design, and Quantitative or Qualitative Methods. Syllabi for identified courses were obtained from

publicly available sources, such as directly from program websites or from databases like OER Commons. To limit the potential for program or staff bias, we did not contact programs to request syllabi. We limited inclusions to documents from 2010 or later to ensure information was up to date. We were able to collect a syllabus for courses from a total of 50 programs.

We conducted text analysis of the syllabi, looking for several key words and phrases that indicated both explicit and implicit inclusion of data management and sharing topics in coursework. Explicit mentions of formal learning included "data management," "data sharing," "data archival," or "data preservation." For nonformal learning, explicit mentions included "secondary analysis," "archive," and "database." We also counted occurrences of additional implicit indicators of data management and sharing topics to identify when such content was likely to be included formally in class or nonformally in assignments, but was not detailed within the syllabi itself. Categorization of keywords and phrases utilized in the analysis is detailed in Table 4.

Table 4. Explicit and implicit words and phrases used for analysis of formal and nonformal syllabi content.

	Explicit Mentions	Implicit Mentions
Formal	Data sharing Data management Data archival/preservation	Research ethics
Nonformal	Secondary analysis Archive Database	Statistics Data collection Data analysis Data presentation/visualization

We searched through the syllabi to highlight the predetermined keywords. We highlighted both exact mentions of phrases like "data collection" and "data analysis," but also highlighted synonymous phrases such as "gathering of data." Additionally, implied use of the word "data," such as in a list of actions applied to data, were also included as multiple unique mentions (e.g. "data collection, management, and analysis" would be considered three mentions of "data"). Lastly, we read syllabi to identify nonformal course requirements of research projects, assignments requiring data collection, or other types of research related presentations.

Syllabi Analysis Results

The syllabi analysis allowed us to further explore the formal and nonformal training opportunities that students experience as they earn their graduate degrees. Similar to the results observed in the survey, syllabi suggested that data management and data sharing are typically not included in formal training for social science graduate programs. Of the 140 programs in the sample, we were able to access syllabi for 50 programs. Most of the programs were doctoral (36, 72%), and a quarter of syllabi (14, 28%) came from master's level programs. Figure 3A shows the breakdown of how many syllabi out of the possible 20 searched for were available, by field of study. Our corpus of syllabi

consisted primarily of documents from political science and economics courses, with the fewest syllabi from history programs.

Although course titles varied from one field to the next, we categorized them into five main areas based on the material included in the syllabus: 1) Research methods courses; 2) qualitative research or data courses; 3) quantitative research or data courses; 4) statistics or analysis courses specific to the discipline; and 5) research design courses. The vast majority of our syllabi fell into either the research methods course category (18, 36%) or the analysis or statistics course category (17, 34%) (see Figure 3B). Two thirds of these courses included nonformal experiences through research projects, specifically papers or presentations that required some form of data analysis or manipulation. These were sometimes replication or secondary analysis projects. Only 30% (15) of these courses required that the students collect new data.

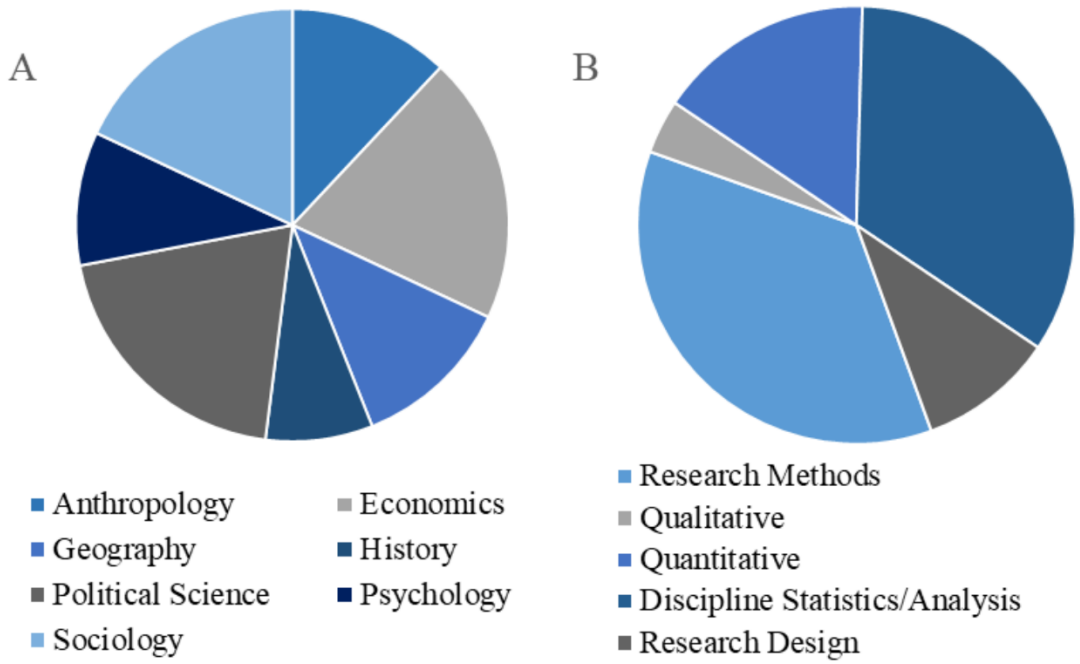


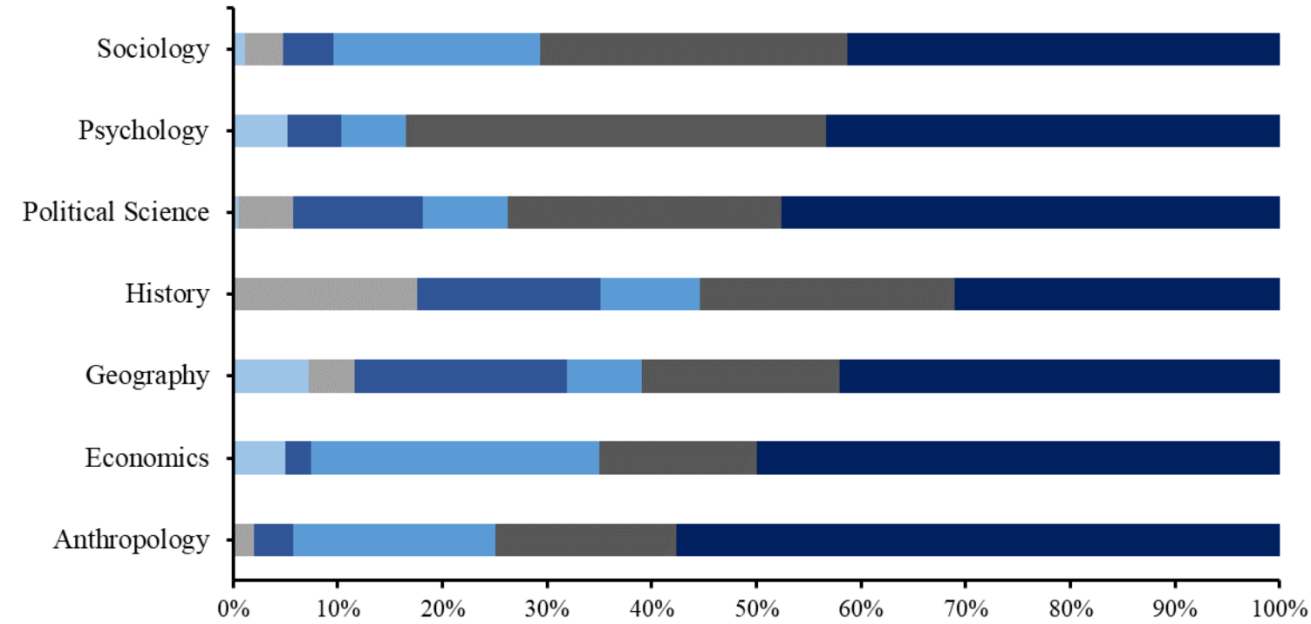
Figure 3. (A) Proportion of fields represented in the final syllabi corpus. (B) Categorization of the type of course syllabi were available from.

When looking at data mentions across all syllabi, we saw a mixture of results (Table 5). Only six of the 50 syllabi did not mention the word "data" at all. Most of the mentions were implicit mentions of nonformal learning experiences, focused on analysis or how to interpret previously collected data. Explicit mentions of data management were included, but not as frequently as the survey data suggested it should be. Given that 75% (18) of the programs in the original survey reported that data management was included within their research methods courses, we expected to see significantly higher mentions of data management in research methods syllabi, especially explicit mentions. As anticipated from previous trends in the survey, however, data sharing appeared very infrequently. On average, it was mentioned significantly less than once per syllabus, with the maximum number of data sharing mentions in a single syllabus being four times and 43 (86%) syllabi not mentioning data sharing at all.

Table 5. Average keyword mentions by learning and mention type.

	Mention Type	Keywords	Mean	SD
Formal Learning	**Explicit**	Data sharing	.30	.839
		Data archival/preservation	.68	1.720
		Data management	1.38	2.465
	Implicit	Research ethics	1.98	3.236
Nonformal Learning	**Explicit**	Archive	.44	1.643
		Database	.92	1.988
		Secondary analysis	1.22	2.410
	Implicit	Data collection	1.78	2.698
		Data presentation/visualization	3.78	3.616
		Data analysis	6.26	6.863
		Statistics	9.44	10.363

The graph in Figure 4 represents the distribution of average “data” keyword mentions for each field. It illustrates the clear lack of inclusion of data sharing, as four of the seven fields have no to almost no mentions of data sharing. There are gaps in data management inclusion as well – a startling prospect given that nearly a third of the courses required students to collect data. This exclusion of explicit data management training in syllabi could suggest that the material is absent from formal coursework. The majority of data mentions trended towards implicit mentions of nonformal learning, and significantly less mentions were of implicit formal education on either data management or data sharing.



	Anthropology	Economics	Geography	History	Political Science	Psychology	Sociology
Data Sharing	0	0.2	0.83	0	0.1	1	0.22
Data Archival	0.17	0	0.5	3.25	1.1	0	0.67
Data Management	0.33	0.1	2.33	3.25	2.6	1	0.89
Data Collection	1.67	1.1	0.83	1.75	1.7	1.2	3.67
Data Presentation	1.5	0.6	2.17	4.5	5.5	7.8	5.44
Data Analysis	5	2	4.83	5.75	10	8.4	7.67

Figure 4. The relative proportion of the total "data" mentions in each category is graphed by discipline to indicate average use within syllabi from a field. The table included in figure 4 shows the breakdown of the mean number of mentions in each data category by syllabi discipline.

Looking across the syllabi to any reference or mention of some archive, database, or academic library, we saw an inclusion rate of 22%. Ethics appeared regularly in the syllabi, although usually mentioned in passing. Most references to ethics were topics for discussion or IRB instructions. The presence of ethics as a topic for course discussion suggests that the ethical code of the governing body, or at least standard ethical practices of research, is discussed in class even when not detailed in the syllabi. In contrast, the majority of content in the syllabi we examined was focused on statistics, both in learning methodology and also in applying statistical analyses to data. On average, more time was devoted to the discussion and inclusion of statistics than to important aspects of ethics. Even when examining these implicit mentions of formal versus nonformal learning, syllabi tended to more heavily focus on nonformal types of learning than on formal learning experiences.

Discussion

Conclusions

The survey and syllabi studies suggest a lack of both formal and nonformal training on data management and data sharing. While our survey suggested that data management content would be represented in research methods courses, references in syllabi appeared only in limited amounts. Training in data sharing or data repository use also appears to be very limited. Our survey showed limited inclusions of data sharing content, even informally, and syllabi contained minimal references to archives or databases, even as resources. In contrast, syllabi indicate that statistics and data analysis are heavily covered, but how to care for data and comply with ethical requirements was often overlooked. Data management and data sharing were rarely included in the overview of courses provided in a syllabus. Our results suggest that despite a strong level of formal training for collecting and analyzing data, minimal effort is spent on training graduates to make data reusable, or on what should become of the data after a project is completed.

Implications

Early education and training in data management and sharing could help alleviate the skills gap. If graduate students' education included formal instruction in data management and data sharing, they would likely be much more prepared to archive and preserve their data long-term. What can seem like a daunting task without the proper resources or information, could be made much simpler with adequate and early formal training. Providing formal education on data management and data sharing would allow future researchers to gain practical knowledge in a low-pressure environment with more time to practice and improve skills before applying them in a professional setting. Moving away from the self-taught models and optional supplemental opportunities and focusing on formal training would allow students to begin gaining mastery of important data management skills, simplifying their future research careers.

Based on our findings, we suggest that data service professionals continue to help fill the existing gaps in data management and sharing training, but that these efforts should be focused on formal and nonformal training options. Until it becomes regular practice for graduate programs to incorporate formal training, data service professionals should consider focusing efforts on providing some supplemental, nonformal coursework or workshops in data management. This could look similar to the courses offered by the ICPSR Summer Program, which provide quantitative methods training for the Social Sciences in workshops and four- and eight-week class sessions. Another potential avenue for this is partnering with graduate programs to offer certifications for graduate students. Through the completion of supplemental coursework, graduate students could walk away with well-developed skills and proof from reputable organizations that they received formal training. In the future, data service professionals could form partnerships with graduate programs to help provide formal coursework – either aiding instructors in developing discipline-specific courses, or delivering these courses themselves.

Limitations

The current research faces several limitations, most stemming from the low survey response rate. Potential survey respondents reported not feeling comfortable with responding to the survey for two main reasons. First, some felt their program did not fit into the description of social science, which could account for the low response rates for fields like history and geography, which are sometimes classified as part of the humanities and physical sciences, respectively. Second, questions from a small number of potential respondents suggested that individuals in the sampling pool felt that because their program did not provide information on data management or data sharing, the survey did not apply to them. Ultimately, the small sample size in both the survey and syllabi study limited cross field comparisons. When conducting the syllabus analysis, it was often difficult to gain access to syllabi, and at times syllabi found were older than our cut-off criteria of 2010 and could not be considered valid. Future analyses may require directly asking course instructors for copies of their current syllabi. Additionally, the lack of syllabi formatting consistency made it difficult to make contextual analyses of where data mentions were appearing in the text. Weighting could not be assigned to item locations because syllabi were not arranged or subdivided in a consistent way. While the syllabi analysis allows an unbiased insight into the potential content of instructors' courses, it is also limited by the fact that syllabi do not necessarily capture all that included in a given class.

Future Directions

The current exploratory analysis highlights the need for additional investigation into graduate level data management and data sharing training. An expansion of our original survey, drawing a larger sample from more social science fields and with an international scope, would help to more clearly define the current approaches. After social science education practices in data management and sharing have been more fully explored, more work could be done to conduct cross disciplinary analyses. Exploring both between social science fields and between the social and physical sciences would allow for a better model for training researchers on data management and data sharing skills. Such comparisons could help identify promising supplemental trainings or modified graduate training strategies.

Additionally, a survey of graduate students would help reveal the supplemental training and experiences that students seek out, as well as their confidence in managing data. A longitudinal survey of graduate students at the beginning and middle of their programs, and then again after graduation or as they enter their careers could help us better understand the gaps in data sharing and management education so that we can better help student researchers become successful career scientists. Overall, our findings suggest a deficit in data management and data sharing education, and we welcome future research to further explore how this can be overcome.

Acknowledgements

This work would not have been possible without the financial support of the Interuniversity Consortium for Political and Social Research. The authors thank Jai Holt for aiding in the development of the graduate program survey, useful discussion, and

initial project design. We also thank Dr. Peter Granda for supervision of the research group and to Kathryn Lavender for administrative support for the beginning the project.

References

[web page] American Economic Association. (2018). AEA Code of Professional Conduct. Retrieved from https://www.aeaweb.org/about-aea/code-of-conduct

[journal article] Bishoff, C., & Johnston, L. (2015). Approaches to Data Sharing: An Analysis of NSF Data Management Plans from a Large Research University. *Journal of Librarianship and Scholarly Communication, 3*(2), 1–27. https://doi.org/10.7710/2162-3309.1231

[journal article] Carlson, J., Johnston, L., Westra, B., & Nichols, M. (2013). Developing an Approach for Data Management Education: A Report from the Data Information Literacy Project. *International Journal of Digital Curation, 8*(1), 204–217. https://doi.org/10.2218/ijdc.v8i1.254

[article] Carlson, J., & Stowell Bracke, M. (2015). Planting the Seeds for Data Literacy: Lessons Learned from a Student-Centered Education Program. *International Journal of Digital Curation, 10*(1), 95–110. https://doi.org/10.2218/ijdc.v10i1.348

[report] Coombs, P. H., & Ahmed, M. (1974). *Attacking Rural Poverty How Nonformal Education Can Help.* Washington, DC.

[journal article] Dikovic, M., & Plavsic, M. (2015). Formal Education, Non-formal and Informal Learning: Knowledge and Experience. *Progress: Journal of Pedagogical Theory and Practice, 156*(1–2), 9–24.

[datasets] Doonan, A. (2019). *Data Management and Data Sharing Training in Social Science Graduate Programs, United States, 2017-2018* (openICPSR-v1) [data files]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E110241V1

[journal article] Hou, C.-Y., Soyka, H., Hutchison, V., Serna, I., Allen, C., & Budden, A. (2017). Evaluating the Effectiveness of Data Management Training: DataONE's Survey Instrument. *International Journal of Digital Curation, 12*(2), 47–60. https://doi.org/10.2218/ijdc.v12i2.508

[book section] Jahnke, L. M., & Asher, A. (2012). The Problem of Data: Data Management and Curation Practices among University Researchers. In *The Problem of Data* (pp. 3–31). Washington, DC: Council on Library and Information Resrouces. Retrieved from http://www.clir.org/pubs/reports/pub154

[journal article] Kelly, K., Marlino, M., Mayernik, M. S., Allard, S., Tenopir, C., Palmer, C. L., & Varvel Jr., V. E. (2013). Model Development for Scientific Data Curation Education. *International Journal of Digital Curation, 8*(1), 255–264. https://doi.org/10.2218/ijdc.v8i1.258

[journal article] La Belle, T. J. (1982). Formal, Nonformal and Informal Education: A Holistic Perspective on Lifelong Learning. *International Review of Education, 28*(2), 159–175. Retrieved from https://www.jstor.org/stable/3443930

[journal article] Lohman, M. C. (2005). A survey of factors influencing the engagement of information technology professionals in informal learning activities. *Human Resource Development Quarterly, 16*(4), 501–527. https://doi.org/10.1002/hrdq.1153

[journal article] Savage, C. J., & Vickers, A. J. (2009). Empirical study of data sharing by authors publishing in PLoS journals. *PLoS ONE, 4*(9), 9–11. https://doi.org/10.1371/journal.pone.0007078

[journal article] Tenopir, C., Allard, S., Sinha, P., Pollock, D., Newman, J., Dalton, E., … Baird, L. (2016). Data Management Education from the Perspective of Science Educators. *International Journal of Digital Curation, 11*(1), 232–251. https://doi.org/10.2218/ijdc.v11i1.389

[web page] Zacharias, M. C. (2010). Scientists Seeking NSF Funding Will Soon Be Required to Submit Data Management Plans. Retrieved April 14, 2019, from https://www.nsf.gov/news/news_summ.jsp?cntn_id=116928