Drowning in Research Data: Addressing Data Management Literacy of Graduate Students

Lise Doucette and Bruce Fyfe

Introduction

Graduate students work in increasingly complex research environments where advances in technology and research methodologies result in gathering and analyzing large amounts of data. Proper management (organization, protection, preservation, sharing) of this research data is essential for productivity, securing grant funding, enabling collaboration and ensuring the future use of data. There are numerous studies of faculty researchers and their research data management practices but relatively few on graduate students, who are also key members of research teams. In particular, little has been studied on how graduate students learn about research data management and how that relates to their research behaviours.

There are many benefits associated with effective Research Data Management (RDM) practices. First and foremost, the advancement of scholarly research is premised on building on a reliable and complete record of previous research, "including the portion in digital form."[1] Creation of a RDM plan that allows for the re-use of publicly funded data is a necessary condition in many granting agency funding requests.[2, 3] Best practices in RDM protect the enormous financial and time investments that have been made by mitigating data loss and avoiding the need for duplication of efforts to recreate lost data.[4] Access to effectively managed research data allows other researchers to validate existing research and to create new knowledge by accessing and building on the work of others.

In this paper we will discuss findings from our research study of social sciences and science graduate students' levels of research data management literacy, which include attitudes and behaviours, and formal and informal education experiences. Using an online survey of Canadian graduate students in the social sciences and science, we were able to reach a large number of students across the country and to gather sufficient responses to allow us to offer some insights on the overall graduate student research data management landscape.

Definitions

Digital Research Data: There are numerous context-specific definitions of research data. For the purposes of this study the definition of research data has been taken from a *Canadian Social Science and Humanities Research Council* (SSHRC) needs assessment survey: research data includes "digital information that has been structured by methodology for the purpose of producing new knowledge."[5] The content of digital data collections may include text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc., "generated by various means including observation, computation, or experiment."[6]

Research Data Management: The life-cycle model of research data consists of data production, data dissemination, long-term management and, discovery and repurposing of data,[7] requiring good research data management practices. RDM is most commonly

Lise Doucette is Research & Instructional Librarian in the Allyn & Betty Taylor Library at Western University, e-mail: ldoucet@uwo.ca; Bruce Fyfe is Research & Instructional Librarian at The D.B. Weldon Library at Western University, e-mail: bfyfe@uwo.ca

defined as activities that fall outside of the work of creation and analysis of data. For this study, RDM refers to the activities involving the organization, protection, preservation and sharing or distribution of the data. Examples of data management practices include: organizing data (naming conventions, version control); providing documentation for or descriptions of data; backing up data; storing data; ensuring the confidentiality of data; and providing access to data.

Literature Review

Studies of research faculty behaviour, competencies and attitudes related to RDM practices provide useful background information and context for our study on graduate students. A number of survey, case-based, interview and focus group studies have been conducted on the data management practices of faculty researchers and operational units within universities. Surveys have been administered to determine the types of data sets created by faculty researchers, how they are managed and the attitudes toward these management activities,[8] helping the researchers identify gaps in best practices and receptivity to educational opportunities.[9] Surveys and case studies have identified key data management gaps in the life-cycle of research data[10] and the risk factors that might limit longer term access to and re-use of research data.[11] A focus group study of research faculty identified barriers to the use and management of research data in a clinical setting[12] while interview-based studies of faculty members sought to develop a comprehensive understanding of research data management workflows, infrastructure in specific research settings, barriers to effective data curation, the existence of gaps in user needs and general attitudes surrounding the sharing of data.[13,14] However, these results are not necessarily generalizable to graduate student researchers, in particular because graduate students are junior researchers newly immersed in the research environment.

Several studies directly examine RDM in a graduate student context, but most are from the perspective of how the library can support their local researchers. For example, one early study using focus groups and interviews identified attitudes and opportunities for libraries related to data organization, security, and the preservation and sharing of data.[15] Using a series of semi-structured interviews of doctoral students, one study sought to build RDM capacity by identifying practical support for researchers.[16] One study, examining post-graduate research behaviour in the Australian context, focused on two broad-based surveys of capabilities and skills within institutions, seeking to understand current practices and training requirements.[17]

Additionally, some universities, research faculties and academic libraries have recognized the need to address research data management education and have responded by developing formal learning opportunities; graduate and undergraduate curriculum promoting best practices for the preservation of scientific data,[18] graduate level business courses on data management issues,[19] and graduate skills workshops[20] have been offered.

This paper contributes to the literature by offering a large scale multi-institution study of differences across faculties and degree level in research-intensive Canadian universities, and examines the relationship between informal and formal educational experiences and the research data management behaviours of graduate students.

Methods

We conducted an online survey using Fluid Surveys that consists of 30 questions regarding graduate students' own behaviour, attitudes, and education related to managing research data. Eighteen of the questions are closed-ended in order to facilitate responding and analysis, with options for 'Other' presented where applicable. Six of the questions are Likert-scale questions on a scale of Strongly Agree to Strongly Disagree, with Not Applicable provided. Three of the questions are open ended text boxes to obtain additional descriptive information from the participants. Basic demographic information is collected (type of student, institution, type of research data).

Our sample consists of nine (9) English-language research universities, a geographically stratified sample from the divisions of the Canadian Association of Research Libraries (CARL)—Western, Ontario, Quebec, and Atlantic. As of October 9th, CARL consisted of 25 English-language universities and four French-language universities. The universities in our sample are: University of Alberta, University of Calgary, McGill University, McMaster University, Memorial University, University of Guelph, University of Toronto, University of Victoria, and University of Waterloo. For each university, our population consists of masters and doctoral students in our six chosen subject areas: Geography,

Psychology, Sociology, Chemistry, Physics, and Earth Sciences. Note that department names among universities are variable, and we have chosen the departments that most closely represent our six subject areas. The departments were chosen to provide a broad sciences and social sciences representation and to maintain a manageable workload for the researchers.

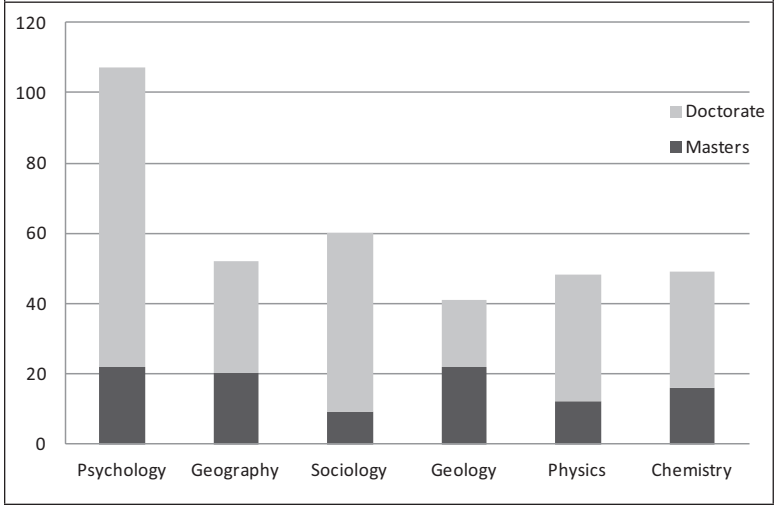
We gathered publicly-available student emails from university and department or research websites. In three cases, we were unable to gather emails for the majority of departments from a chosen university, and we therefore omitted that university and randomly selected one of the remaining universities from that geographic area to study.

An invitation was emailed to 3,503 students on November 28, 2012, providing information for them about the study and asking them to participate in the survey. A follow-up email with much of the same text was sent as a reminder on December 5, 2012. The survey was closed on December 12, 2012.

Results

Our total number of responses was 477; we excluded from our analysis 117 respondents who opened the survey and answered only basic demographic questions. We analysed the remaining 360 responses in SPSS using descriptive and inferential statistics. With N=360, our response rate was 10.3%, and varied across departments, with social sciences students responding at a proportionally higher rate. The breakdown of respondents by department and degree type can be seen in Figure 1.

FIGURE 1
Respondents by Department and Degree Type



For some of our analysis we collapsed the departments into two faculties: Social Sciences (Psychology, Geography, and Sociology) and Science (Geology, Physics, and Chemistry). We chose these assignments of departments to faculties based on the academic departments and faculties at our own university (Western Ontario).

Results presented here reflect a preliminary analysis of the data, and focus on two key areas: respondents' attitudes and behaviours towards research data management (RDM), and literacy and education related to RDM (self-assessment, formal and informal education).

Attitudes and Behaviours Related to Research Data Management

We asked several questions in our survey to determine respondents' attitudes towards RDM, as well as their RDM-related behaviours.

83.2% of respondents (N=310) agreed or strongly agreed with the statement "The management of research data is important for my research group," and 90.0% agreed or strongly agreed with the statement "I am confident in my ability to manage research data."

In subsequent more detailed questions, respondents do indicate that they have experienced situations where poor management of research data has occured. 14.2% of respondents (N=360) indicated they had "re-collected data that you know had been previously collected because you could not find or open the file"; 17.2% (N=360) indicated they had "lost a file and been unable to re-collect the data." In both cases, the majority of respondents who had re-collected data or lost a file also agreed or strongly agreed with the initial statements related to importance (81.1%) and confidence in abilities (76.6%) related to RDM.

Overall, 40.3% of students (N=360) were unsure, disagreed, or strongly disagreed with the statement "I have provided enough documentation that a research peer or future grad student could use my data." Using a t-test to compare means of responses to that statement (where Strongly Disagree = 1 and Strongly Agree = 5), we found no statistically significant difference between faculties (t=1.155, p=0.129) or between degree programs (t=0.512, p=0.609).

73.8% (N=321) respondents agreed or strongly agreed with the statement "For my research there is value in reusing or repurposing research data"; however, 38.0% of those respondents were unsure, disagreed, or strongly disagreed about the statement related to providing sufficient documentation.

37.8% (N=360) of respondents have neither written nor verbal policies related to research data management within their research group.

Literacy and Education

We noted that 90% of respondents agree with the statement "I am confident in my ability to manage research data," a measure of self-reporting RDM literacy. Using a t-test to compare means of responses to that statement (where Strongly Disagree = 1 and Strongly Agree = 5), we found that social sciences students have a significantly higher mean than science students (t=2.799, p=0.005), and that doctoral students have a significantly higher mean that masters students (t=2.597, p=0.010).

Formal Education

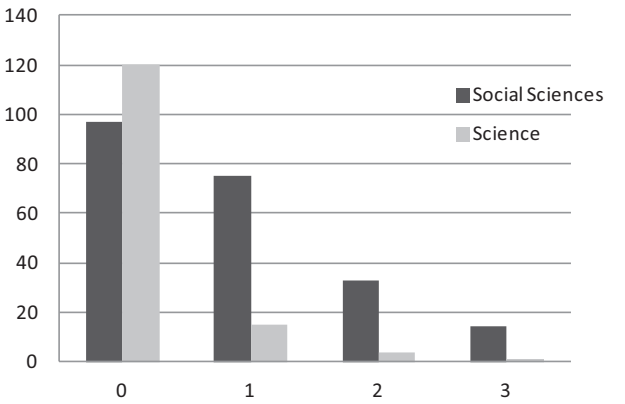
Respondents were asked three questions that indicated the level to which they engaged in formal educational experiences where they learned about managing their research data (N=360 for each of the following):

- 20.8% of respondents took a research methods course in which RDM was discussed
- 22.3% took another course where RDM was discussed
- 15.4% participated in a workshop where RDM was discussed

A "formal education score" was calculated for each respondent, with one point assigned for each 'Yes' or positive response to the three questions above, for up to 3 points total. For example, a student who took a workshop related to RDM and a research methods course in which RDM was discussed would have a score of 2. The range of points from 0-3 is shown below for the faculties of social sciences and science.

Using a t-test to compare means of formal education scores, we found that social sciences students have a significantly higher mean than science students (t=8.701; p=0.000). Using a t-test to compare means of formal education scores, we found no significant difference between doctorate and masters' students (t=1.786; p=0.107).

FIGURE 2
Formal Education Score by Faculty



Informal Education

Respondents were asked three questions that indicated the level to which they informally sought to learn about managing their research data (N=360 for each of the following):

56.1% of respondents agreed or strongly agreed with the statement "I educate myself on best practices for preserving my data"

33.3% of respondents indicated that they engage in self-directed learning related to research data management, and provided textual description

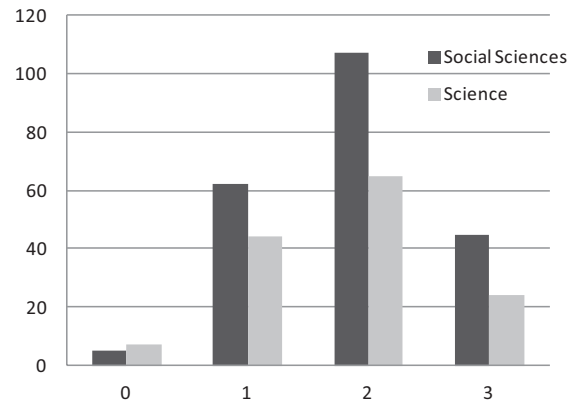
TABLE 1
Number of Respondents Who Have Discussed Research Data Management with the Following

Research faculty member who is my supervisor	249
PhD student	195
Research Faculty Member who is not my supervisor	108
Masters student	108
Undergraduate student	52
Researcher from another institution	43
Post-doctoral fellow	39
Lab technician	37
IT staff	37
Research Office staff	21
Librarian	17
Visiting researcher/scholar	9
Emeritus faculty member	4
Researcher from private industry	4

93.3% of respondents indicated that they have discussed RDM with at least one of the following types of colleagues from the list, with 25 providing additional responses, such as friends/family and external agencies. Table 1 below shows the range of types of colleagues with whom the respondents discussed RDM.

Similar to the "formal education score" calculated above, an "informal education score" was calculated for each respondent, based on a 'Yes' or positive response to the three questions above, for up to 3 points total. For example, a student who indicated that they engaged in self-directed learning and that they spoke with any number of colleagues from the list would have a score of 2. The range of points from 0-3 is shown below for the faculties of science and social sciences.

FIGURE 3
Informal Education Score by Faculty



Using a t-test to compare means of formal education scores, we found no significant difference (t=1.436; p=0.152) between social sciences and science students. Similarly, using a t-test to compare means of formal education scores, we found no significant difference between doctorate and masters' students (t=0.512; p=0.609).

Discussion and Conclusions

As presented in the Results, we noted that students rate both their own confidence/ability and the importance of research data management highly, which, if taken on their own, could seem to indicate that 'all is well.' However, students then demonstrate that some of their practices are in fact indications of poor management of research data that can cause significant reduction in research productivity. For example, the students who recollected data (14.2%) were duplicating either their own efforts or those of a research colleague, potentially at a significant money and time cost. The students who lost a file (17.2%) and could not recollect data permanently lost some valuable information related to their research. 38.0% of students see value in reusing their data but do not feel confident that they have provided enough documentation that a colleague could use their data; their data is much less likely to be able to be reused, therefore preventing future research and possibly also wasting time/money by future researchers who must recollect data. The 37.8% of respondents who have no written or verbal policies related to RDM could have less knowledge and awareness of best practices, leading to lack of proactive management of their research data for their own benefit and for the benefit of research colleagues.

In terms of students' confidence, we found that social sciences students and doctoral students have statistically higher reported levels than do science and masters students, respectively. We expected that doctoral students would be more confident due to their increased experience as researchers. There was no statistically significant difference between degree programs or between faculties for responses to "I have provided enough documentation that a research peer or future grad student could use my data." We expected that doctoral students would more strongly agree with this statement due to their increased experience as researchers.

We found the social sciences students have a statistically higher formal education score than do science students. Research methods courses are more common in the social sciences, and this contributed to the overall higher score for social science students. There was no statistically significant difference between faculty and informal education score; degree and formal education score; or degree and informal education score.

We discovered that very few students (4.7%, N=360) have discussed the management of research data with a librarian. The colleagues they most commonly discuss the topic with are research faculty and other graduate students. Librarians and libraries are not generally part of their research environment or the colleagues they consult on this topic.

Further Research

Students rate both their own confidence/ability and the importance of research data management highly,

and then show contradictory behaviours. Further research could compare these findings to other, more established forms of literacy such as information literacy, a topic that has been extensively studied. There may be lessons learned from information literacy that could apply to research data literacy.

We expected that doctoral students would agree with "I have provided enough documentation that a research peer or future grad student could use my data" at a significantly higher level than masters students, as most of them would have a previous masters' degree and more extensive research experience. This was not the case. This aspect could be further explored, in particular by analyzing textual answers related to how students name their files and variables, or through interviews or focus groups.

Future analysis of 120 textual responses will provide more detail on specific ways that graduate students have undertaken self-directed learning related to RDM. Additionally, analysis of differences in composition among social sciences and sciences research groups, as well as the types of colleagues consulted by graduate students in each of the faculties, will be undertaken.

Notes

1. Beagrie, Neil and Julia Chruszcz and Brian Lavoie. "Keeping Research Data Safe: A Cost Model and Guidance for UK Universities." JISC. April 2008. http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf.
2. National Science Foundation. "Dissemination and Sharing of Research Results," September 7, 2012. http://www.nsf.gov/bfa/dias/policy/dmp.jsp.
3. Social Sciences and Humanities Research Council of Canada (SSHRC). "National Research Data Archive Consultation Phase One: Needs Assessment Report." Library and Archives Canada, 2001. www.sshrc-crsh.gc.ca/about-au_sujet/publications/da_phase1_e.pdf .
4. Shearer, Kathleen and Canadian Association of Research Libraries. "Research Data: Unseen Opportunities." 2010. http://carl-abrc.ca.proxy2.lib.uwo.ca/uploads/pdfs/data_brochure-e.pdf.
5. SSHRC, National Research Data Archive Consultation.
6. National Science Foundation. Dissemination and Sharing.
7. Research Data Canada. Research Data Strategy Working Group. "Stewardship of Research Data in Canada: A Gap Analysis." October 2008. http://rds-sdr.cisti-icist.nrc-cnrc.gc.ca/eng/about/achievements.html.
8. Parham, S. W., J. Bodnar, and S. Fuchs. "Supporting tomorrow's Research Assessing Faculty Data Curation Needs at Georgia Tech." College & Research Libraries News 73, no. 1. 2012: 10-13. http://crlnews.highwire.org/content/73/1/10.full.
9. Scaramozzino, JM, ML Ramirez, and KJ McGaughey. "A Study of Faculty Data Curation Behaviors and Attitudes at a Teaching-Centered University." College & Research Libraries 73, no. 4. 2012: 349-365. http://crl.acrl.org/content/early/2011/08/26/crl-255.full.pdf+html.
10. O'Reilly, Kelley, Jeffrey Johnson, and Georgiann Sanborn3. "Improving University Research Value A Case Study." Sage Open 2012. http://sgo.sagepub.com/content/2/3/2158244012452576.
11. Knight, G. "A Digital Curate's Egg: A Risk Management Approach to Enhancing Data Management Practices." Journal of Web Librarianship 6, no. 4. 2012: 228-250. http://resolver.scholarsportal.info.proxy1.lib.uwo.ca/resolve/19322909/v06i0004/228_adcearatedmp.
12. Bardyn, Tania P., Taryn Resnick, and Susan K. Camina. "Translational Researchers' Perceptions of Data Management Practices and Data Curation Needs: Findings from a Focus Group in an Academic Health Sciences Library." Journal of Web Librarianship. 6, no. 4. 2012: 274-287. http://resolver.scholarsportal.info.proxy1.lib.uwo.ca/resolve/19322909/v06i0004/274_trpodmiaahsl.
13. Jahnke, Lori M. and Andrew Asher. "The Problem of Data: Data Management and Curation Practices among University Researchers." Council on Library and Information Resources. 2012. http://www.clir.org/pubs/reports/pub154/problem-of-data.
14. Diekmann, F. "Data Practices of Agricultural Scientists: Results from an Exploratory Study." Journal of Agricultural & Food Information 13, no. 1 (2012): 14-34. http://resolver.scholarsportal.info.proxy1.lib.uwo.ca/resolve/10496505/v13i0001/14_dpoasrfaes.
15. Marcus, C., S. Ball, L. Delserone, A. Hribar, and W. Loftus. "Understanding Research Behaviors, Information Resources, and Service Needs of Scientists and Graduate Students: A Study for the University of Minnesota Libraries." (2007). http://purl.umn.edu/5546.
16. Ward, C., L. Freiman, L. Molloy, S. Jones, and K. Snow. "Making Sense: Talking Data Management with Researchers." International Journal of Digital Curation 6, no. 2 (2011). http://www.ijdc.net/index.php/ijdc/article/view/197.
17. Henty, Margaret and Weaver, Belinda and Bradbury, Stephanie J. and Porter, Simon "Investigating Data Management Practices in Australian Universities." (2008). http://eprints.qut.edu.au/14549.
18. Piorun, M. E., D. Kafel, T. Leger-Hornby, S. Najafi, E. R.

Martin, P. Colombo, and N. R. LaPelle. "Teaching Research Data Management: An Undergraduate/Graduate Curriculum." Journal of eScience Librarianship 1, no. 1 (2012): 8.

19. University of Calgary, MGIS 737—enterprise data management| Haskayne School of Business. 2011. http://haskayne.ucalgary.ca/courses/w11/MGIS737?destination=profiles/183-44619/courses.
20. Carleton University Library. 2012. "Professional Skill Workshops for Graduate Students—Introduction to Data Management." Sept 27, 2012. http://www.library.carleton.ca/library-news/professional-skill-workshops-graduate-students-introduction-data-management .