Research Data Management Librarian Academy Project Report

Compiled by Alyson Gamble, Data Fellow

28 September 2018

Introduction

Summary

Research data management (RDM) operates on the intersections of data, information, and library sciences. The Research Data Management Librarian Academy, or RDMLA, is a group of several university libraries partnering with Elsevier to study the need for a Research Data Management Librarian Academy that could provide online RDM training. The team has worked to compile an inventory of existing courses for academic librarians as well as researchers and conduct a needs assessment through interviews, surveys, and focus groups in order to identify gaps in current training offerings and to identify what librarians and researchers need to contribute to their success. The project partners consider the project essential to gain efficiencies, meet growing RDM needs, and address the librarian skills gap.

Partners

The partners on this project include: Boston University, Elsevier, Harvard Library, Harvard Medical School, Massachusetts College of Pharmacy and Health Science University, Northeastern University, Simmons University, and Tufts University.

Results

Survey Results

The Research Data Management Librarian Academy (RDMLA) Survey was intended to gauge general perceptions, particularly from librarians, about RDM and RDM training. The RDMLA survey was launched via SurveyMonkey on May 23, 2018 and closed on September 4, 2018. It contained 19 questions developed by Elsevier's Jean Shipman and Harvard Medical School's Dr. Elaine Martin. The survey was distributed to relevant groups via email, a Library Connect blog post, listservs, and word of mouth. At the conclusion of the survey on September 4, 2018, there were 241 responses. The majority of these responses (*n* = 116) were collected during May 2018. The 628 open answer responses were hand-coded by Alyson Gamble of Simmons University.

Landscape

A portrait of the RDM landscape was gathered from the RDMLA survey. Sixty-three respondents to a question about the RDM services at their institutions reported that they offer: RDM planning (*n* = 50, 79%); data sharing and dissemination (*n* = 48, 76%); data discovery and access (*n* = 41, 65%); data preservation (*n* = 41, 65%); metadata (*n* = 40, 63%); data

organization and curation (*n* = 36, 57%); data visualization (*n* = 35, 56%); protocol documentation (*n* = 18, 30%); and other RDM services (*n* = 18, 30%). Of these other services, respondents indicated that their institutions offered: RDM training (*n* = 6), consultations (*n* = 4), software access and/or training (*n* = 3), information on best practices (*n* = 2), data analysis (*n* = 2), and preservation (*n* = 1).

Fifty-seven respondents indicated the primary tools used by their institutions. Data repositories (*n* = 44, 77%) and data processing software (*n* = 38, 67%) were the most commonly used tools. Twenty-three respondents indicated their institutions used a data citation manager, 21 indicated use of electronic lab notebooks, 18 indicated use of pre-prints, and 16 indicated use of a data search engine. Thirteen respondents indicated their institution utilized a tool not listed in the survey question; six of these used DMPTool.

Two hundred and thirty-nine respondents rated the status of the current development level of their role development in RDM at their institution. The majority of programs were rated as either not developed (*n* = 89, 37%) or almost not developed (*n* = 58, 24%). Forty-nine programs ranked as neither developed nor not developed, 24 as developed, and 19 as very developed. Regarding their role in RDM at their institutions, 137 respondents indicated they wished they had a more formal role with RDM. Sixty-one respondents felt prepared to offer RDM services, while 26 respondents did not.

Fifty-three respondents indicated several ways they saw librarian roles evolving within the context of RDM: through application of librarian skills that are already developed (*n* = 24, 45%); through infrastructural development and change (*n* = 23, 43%); in ways related to consultations/instruction/training (*n* = 18, 34%); in application to data curation (*n* = 10, 19%); in relation to new technology (*n* = 10, 19%); and in ways related to staffing (*n* = 4, 8%).

Current training landscape

Respondents to the RDMLA survey offered insight into the RDM training they had already received. Thirty-two had received training in data curation, 28 in data management/data management plans, 16 in metadata, 13 in data preservation/repositories/storage, 12 in software/coding/technical skills, 11 in infrastructure issues (including the value of librarians), eight in discipline or funder-specific practices/tasks/tools, eight in data literacy, eight in consultations/instruction/outreach, six in “everything” or an equivalent, and four in data security (including de-identification). Three respondents indicated they had received this training via hands-on activities.

Demonstrated need for training

Fifty-four RDMLA survey respondents who stated they did not feel prepared to offer RDM services indicated why. Of these, 39 stated that knowledge in the area, potentially acquired through training, could help them feel more prepared to offer RDM services. Fifty-four RDMLA survey respondents indicated challenges offering RDM at their institutions beyond lack of training, though 10 of these respondents stated that a lack of training was what challenged the development and implementation of RDM at their institution.

Delivery of training

A majority of RDMLA survey respondents who indicated their likelihood of participating in online RDM training (*n* = 197) stated they would be "very likely" (*n* = 93, 47%) or "somewhat likely" (*n* = 69, 35%) to participate in online RDM training. The 196 RDMLA survey respondents who indicated interest in receiving continuing education (CE) credit for taking online RDM training had a slight preference (*n* = 102, 52.04%) for receiving Continuing Education (CE) credit in exchange for training.

Subjects for training

Twenty-six RDMLA survey respondents who indicated they did not feel personally prepared to offer RDM services described a lack of knowledge/training in several areas. These included: early steps of data/data management (*n* = 2); specific software/data mining (*n* = 2); scientific data (*n* = 2); data preservation/repositories/storage (*n* = 2); general data management (*n* = 1); and policies (*n* = 1). One respondent indicated they needed hands-on training.

Sixty-six RDMLA survey respondents indicated they would like further RDM training for themselves in several different areas. These included software/coding/technical skills (*n* = 23); discipline or funder-specific practices/tasks/tools (*n* = 14); infrastructure issues (*n* = 10); data management/data management plans (*n* = 8); consultations/instruction/outreach (*n* = 7); "everything" or an equivalent (*n* = 6); collection of RDM resources (*n* = 6); data security (including de-identification) (*n* = 5); metadata (*n* = 5); and data repositories/data storage (*n* = 4). Eight respondents indicated they would like hands-on training in RDM.

Sixty-one RDMLA survey respondents indicated LIS professionals needed training in several areas.[1] These included consultations/instruction/outreach (*n* = 14); software/coding/technical skills (*n* = 11); "everything" or an equivalent (*n* = 11); infrastructure issues (*n* = 10); discipline or funder-specific practices/tasks/tools (*n* = 9); data curation (*n* = 4); data repositories/storage (*n* = 4); data management/data management plans (*n* = 2); and data security (including de-identification) (*n* = 2). Three respondents indicated LIS professionals needed further hands-on training in RDM.

Sixty-eight RDMLA survey respondents indicated desired training opportunities for researchers. These training opportunities included data curation (*n* = 22); "everything" (*n* = 15); data management/data management plans (*n* = 14); infrastructure issues (*n* = 12); discipline or funder-specific practices/tasks/tools (*n* = 12); metadata (*n* = 8); value of librarians in RDM (*n* = 7); data security (including de-identification) (*n* = 6); data repositories/data storage (*n* = 6); software/coding/technical skills (*n* = 6); and data literacy (*n* = 2).

From these four questions, a total of 221 responses indicated training was needed in several areas. These included software/coding/technical skills (*n* = 42); infrastructure issues (including value of librarians in RDM) (*n* = 39); discipline or funder-specific practices/tasks/tools (*n* = 38); "everything" (*n* = 34); data curation (*n* = 31); data management/data management plans (*n* = 25); consultations/instruction/outreach (*n* = 21); data

[1] Phrasing of this question elicited 26 comments about the question itself.

preservation/repositories/storage (*n* = 16); metadata (*n* = 13); data security (including de-identification) (*n* = 6); and data literacy (*n* = 2). Twelve of the 221 respondents indicated there was need for hands-on training in RDM.

Collection of RDM Materials

In the RDMLA survey, several respondents (*n* = 10) indicated they, other LIS professionals, or researchers could benefit from a space for collected resources about RDM. This was echoed in other responses to the survey, including the disjointed nature of how people find information about RDM training. Of 125 respondents who indicated how they identify opportunities, 66 found training through word of mouth, 36 found training through conferences, 22 found training through websites, 12 found training through workshops, 11 found training through coursework, nine found training through publications, and eight found training through webinars.

E-mail Survey Results

An e-mail survey was conducted with educators from library and information science (LIS) programs. These educators were recruited via the ALISE listserv and direct outreach to individuals. Seven responses were collected by September 10, 2018. Six schools were located in North America and one in Europe. The e-mail survey was comprised of eight questions with eight subquestions. A copy of this instrument is attached to this report.

All seven respondents reported their school offered courses in RDM. These courses included: data analytics (*n* = 3), data curation (*n* = 3), data visualization (*n* = 3), data mining (*n* = 2), data science (*n* = 2), management (*n* = 2), text mining (*n* = 2), and one each in: data cleaning; data science storytelling, data services; data stewardship; databases; digital libraries; digital preservation; information processing; information retrieval; laws for research data; metadata; project management; research data management; research data infrastructure; statistical modeling; and value of research data. One of the schools offered a separate degree focused in RDM, while the other six did not. Two of the schools required enrollment in a degree program in order to pursue RDM, while five did not. At the schools, RDM courses were offered online (*n* = 2), in-person (*n* = 2), and in both online and in-person formats (*n* = 3). All seven respondents indicated that students would be interested in online modules if they could get course credit. Course refreshment was the responsibility of the faculty members at four schools, part of regular review at two, and not described at one. Respondents were aware of RDM coursework at Indiana (*n* = 2), Syracuse (*n* = 2), Berkeley (*n* = 1), Drexel (*n* = 1), UCL (*n* = 1), and UNC Chapel Hill (*n* = 1).

Respondents indicated where they saw graduates of RDM coursework working. These environments included: academic (*n* = 4), corporate (*n* = 3), government (*n* = 2), large data centers (*n* = 1), medicine (*n* = 1), research libraries (*n* = 1), and scientific organizations (*n* = 2).

Four of the respondents offered recommendations for building online courses: “Focus on developing excellent content in an asynchronous yet dynamic online environment;” “Our students can integrate some courses into their program of study not offered by [our school] if

they are offered by an accredited program for graduate credit;" "In the UK there has been a lot of work undertaken in this sphere and funded by JISC. This has included the production of training materials;" and "It's an excellent way to deliver content. Students find it to allow them to continue their home life and work life, but still gain a credential or degree."

Two respondents offered additional commentary. One was interested in discussing, "How practitioners with the expertise and interest in teaching in this area could become involved in existing academic programs." The other stated, "Research data management does align to archives and records management skillsets but also library and information science given the explicit drive by [government] to push data into the public domain."

There was a high overlap of the e-mail survey results and the environmental scan of courses at ALA-accredited universities and iSchools.

Competencies Scan

Job Listings

A scan of job listings was conducted from June 20 through September 20, 2018. Jobs were identified from the American Library Association (ALA) and Association of College and Research Libraries (ACRL) JobList, the American Zoo & Aquarium Association: Librarians SIG discussion list, Association for Information Science & Technology (ASIS&T) listservs, the Code4Lib jobs site, the International Association of Aquatic and Marine Science Libraries and Information Centres (IAMSLIC) discussion list, the Simmons University JobLine, Twitter posts with the hashtags #datalibs and #medlibs, and other sources. Forty-four jobs were identified, all based in the United States. Limitations of this scan included a lack of focus on international jobs, as well as unscanned sources including: Hiring Librarians, INALG, LibGig, and the SAA job board.

Job listings were hand-coded according to 16 competency topics, outlined in Table 2. The most frequently occurring competencies were consultations/instruction/outreach (*n* = 27), disciplinary/funder requirements (*n* = 26), data curation (*n* = 18), data management plans/planning (*n* = 17), metadata (*n* = 13), data storage (*n* = 11), infrastructure (*n* = 11), and software (*n* = 10). A complete total can be found in Figure 1.

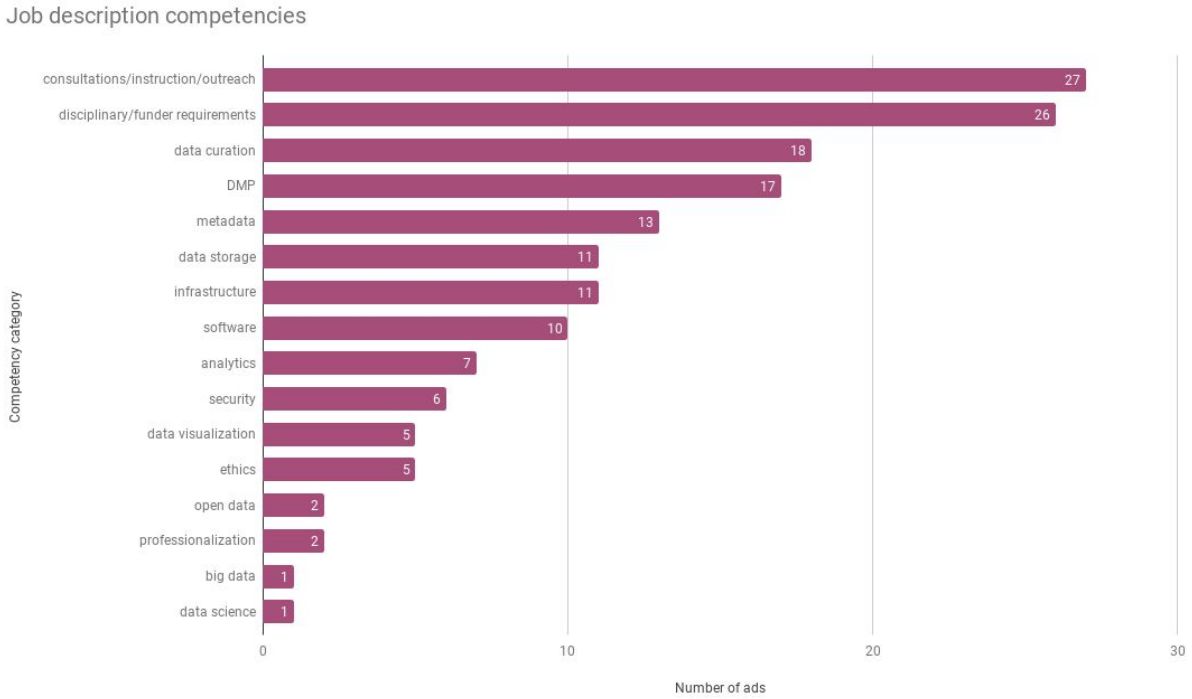


Figure 1. Competencies totaled

European medical librarians' professionalization needs

A discussion of professional competencies on the European Association for Health Information and Libraries discussion list examined the employment of medical librarians. All six replies were from Europe; four came from Italy, one from Spain, and one from the United Kingdom. Responses from Italy and the United Kingdom stated that there is no requirement for medical librarians in these countries. The response from Spain stated that necessary subscriptions to medical journals indicated to most that there was a need for medical librarians. The responses indicated that the competency most needed to improve the perception of medical librarians' worth was professionalization of librarians (*n* = 6), as well as a need for curriculum (*n* = 3) and infrastructure (*n* = 3). A complete total can be found in Figure 2.

Some groups, such as LIBER's Digital Skills for Library Staff & Researchers Working Group and the FOSTER Plus project, are actively working to develop digital skills. In the LIBER and FOSTER Plus discussions, considerations include development of a plan in European Library and Information Science projects. Additionally, at this group's meeting in 2018, the attendees concluded that skills are not confined to single professions.[2]

[2] https://libereurope.eu/blog/2018/07/12/digital-skills-liber-2018/

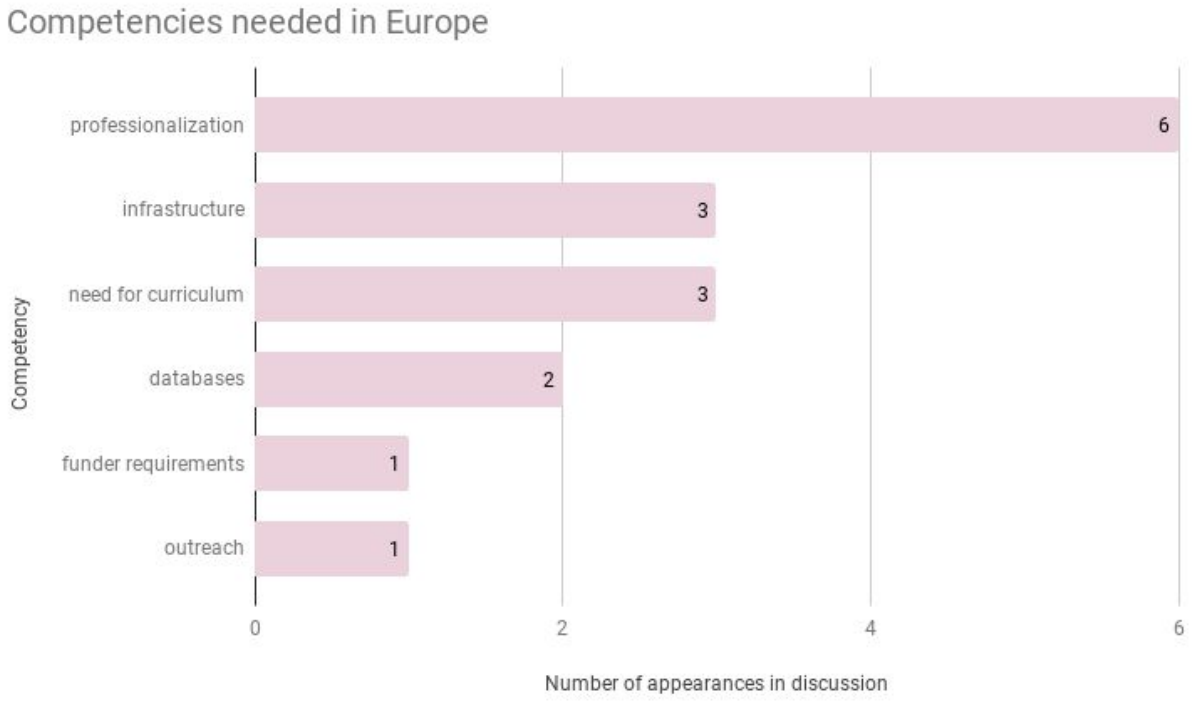


Figure 2. Competencies needed in Europe

Groups and Policies

From examination of 16 websites belonging to 14 different groups from around the world, several different competencies were noted. These included: data curation (*n* = 11), security (*n* = 10), ethics (*n* = 9), data management plans/planning (*n* = 8), consultations/instruction/outreach (*n* = 7), disciplinary/funder requirements (*n* = 7), infrastructure (*n* = 7), data storage (*n* = 6), software (*n* = 4), metadata (*n* = 2), analytics (*n* = 1), data literacy (*n* = 1), data science (*n* = 1), data sharing (*n* = 1), and instruction (*n* = 1). A complete total can be found in Figure 3.

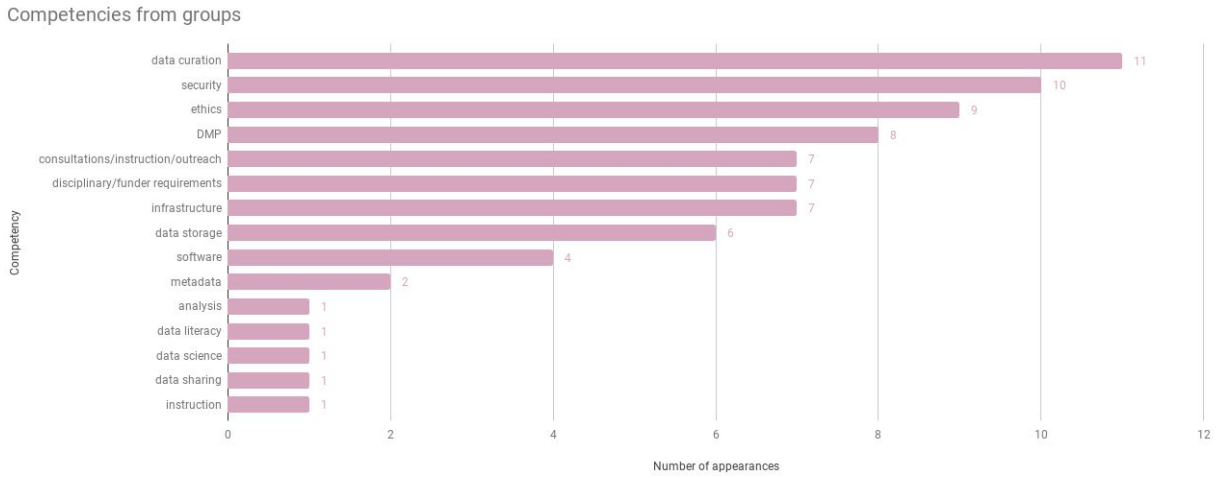


Figure 3. Competencies from group websites

Training

MOOCs, Webinars, and Workshops

One hundred and sixty-four training opportunities were identified via advertisements, listservs, publications, websites, and word-of-mouth. Of the 164 opportunities, 104 were open to anyone, 25 were open to affiliates of the organization offering the training, 11 were open to researchers, ten to LIS professionals, six to conference attendees, and one by invite-only. Seven of the opportunities did not have information about their openness available.

The time commitment for trainings ranged from 45 minutes to one year. The most common training time was one hour (*n* = 24), followed by one day (*n* = 14). Seventy-three opportunities did not list the required time to complete the training. For participants' efforts, six opportunities offered a certificate, six offered a form of continuing education credits, one offered a badge, and 150 did not have information available about participants' documentation.

Training was offered in-person for 54 opportunities, while 44 were offered online. Forty-one offerings were for online materials. Eighteen were for hybrid offerings. Five did not have information available for how the materials were offered.

Seventy-nine trainings were offered for free, while 11 were free for affiliates of the organization offering the training. Paid training costs ranged from 15 USD to 3180 Pound sterling. Two offerings were included solely as part of conference registration. Forty-three offerings did not list their cost.

Topics covered in these training opportunities included data curation (*n* = 52), disciplinary/funder requirements (*n* = 46), DMP (*n* = 45), data storage (*n* = 44), infrastructure (*n* = 36), security (*n* = 33), software (*n* = 32), ethics (*n* = 26), consultations/instruction/outreach (*n* = 23), metadata (*n* = 21), analysis (*n* = 16), data science (*n* = 10), open data (*n* = 10), data visualization (*n* = 9), data literacy (*n* = 6), big data (*n* = 4), and data sharing (*n* = 1). Ten of the opportunities included hands-on elements.

Courses and programs outside of LIS programs

Twenty-two courses or programs from 13 different universities in two iSchool regions were identified that had elements of research data management in their course objectives. Twenty-one courses or programs were currently offered and one was in development. Eleven courses or programs were offered in Europe and 11 were offered in North America. Thirteen of the courses or programs were offered in Master's programs, nine in bachelor's programs, five as graduate certificates, and two in doctoral programs. The course in development did not report the level for the course that would be offered.

Courses or programs covered a variety of topics, including: data curation (*n* = 18), data storage (*n* = 10), analysis (*n* = 7), infrastructure (*n* = 5), ethics (*n* = 4), software (*n* = 4), disciplinary/funder requirements (*n* = 3), consultations/instruction/outreach (*n* = 2), data visualization (*n* = 2), security (*n* = 2), data literacy (*n* = 1), data science (*n* = 1), and metadata (*n* =1).

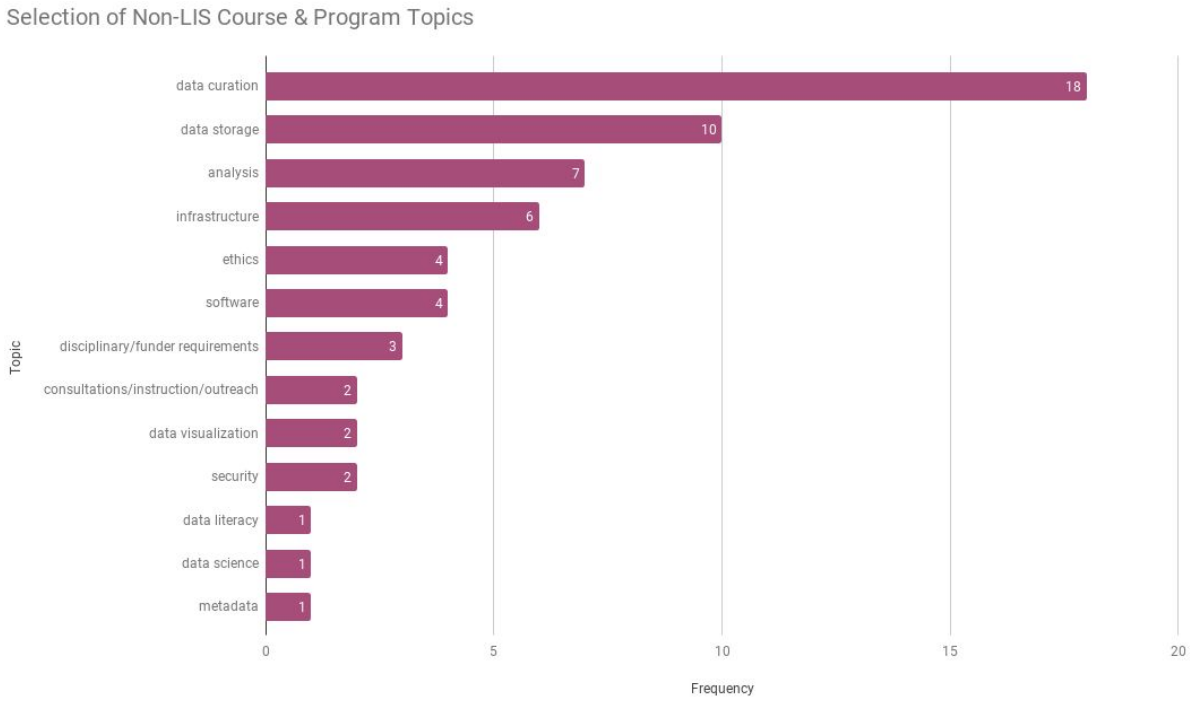


Figure 4. Non-LIS course and program topics

Courses or programs at American Library Association-accredited programs

Twenty-five courses or programs from American Library Association (ALA)-accredited programs at 14 universities in the North America iSchool region were identified that had elements of research data management in their course objectives. Eighteen of the courses or programs

were offered in Master's programs, two in bachelor's programs, six as graduate certificates, and three in doctoral programs.

Courses or programs at the ALA-accredited programs covered a variety of topics, including: disciplinary/funder requirements (*n* = 10), data curation (*n* = 8), software (*n* = 6), data visualization (*n* = 4), ethics (*n* = 4), metadata (*n* = 4), consultations/instruction/outreach (*n* = 3), data science (*n* = 3), data storage (*n* = 2), DMP (*n* = 2), security (*n* = 2), analysis (*n* = 1), big data (*n* = 1), and databases (*n* = 1).

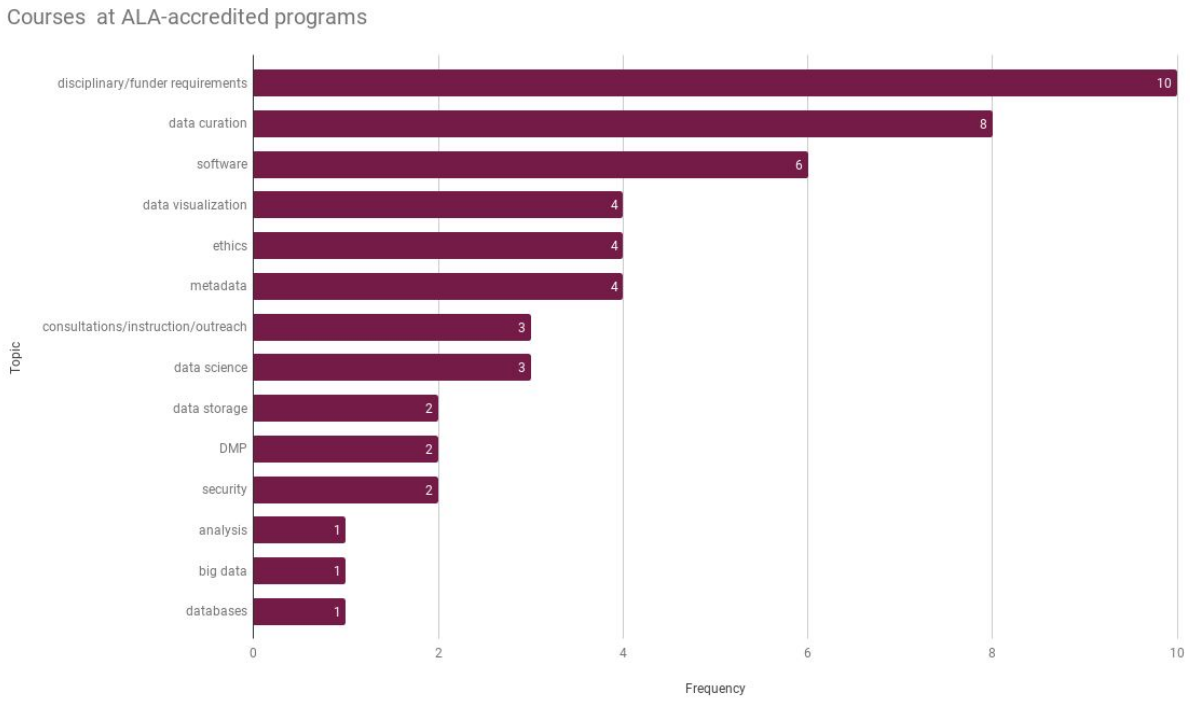


Figure 5. Topics from courses at ALA-accredited programs

Courses and programs at iSchools

One hundred and sixty-five courses or programs from iSchool-affiliated programs at 45 universities that had elements of research data management in their course objectives. Thirty-nine courses or programs were located in North American iSchools, six were located in Asia-Pacific iSchools, and one was located in Europe. Ninety-four of the courses or programs were offered in Master's programs, 56 in bachelor's programs, 23 as graduate certificates, and 16 in doctoral programs. There was no enrollment level information for 28 courses or programs. One was offered open online.

Courses or programs at the iSchool-affiliated programs covered a variety of topics, including: data curation (*n* = 58), disciplinary/funder requirements (*n* = 52), software (*n* = 40), data storage (*n* = 39), analytics (*n* = 37), ethics (*n* = 29), professionalization (*n* = 27), data science (*n* = 25), metadata (*n* = 25), infrastructure (*n* = 24), security (*n* = 20), data visualization

(*n* = 19), databases (*n* = 19), big data (*n* = 17), hands-on (*n* = 11), DMP (*n* = 10), data literacy (*n* = 5), and open data (*n* = 5).

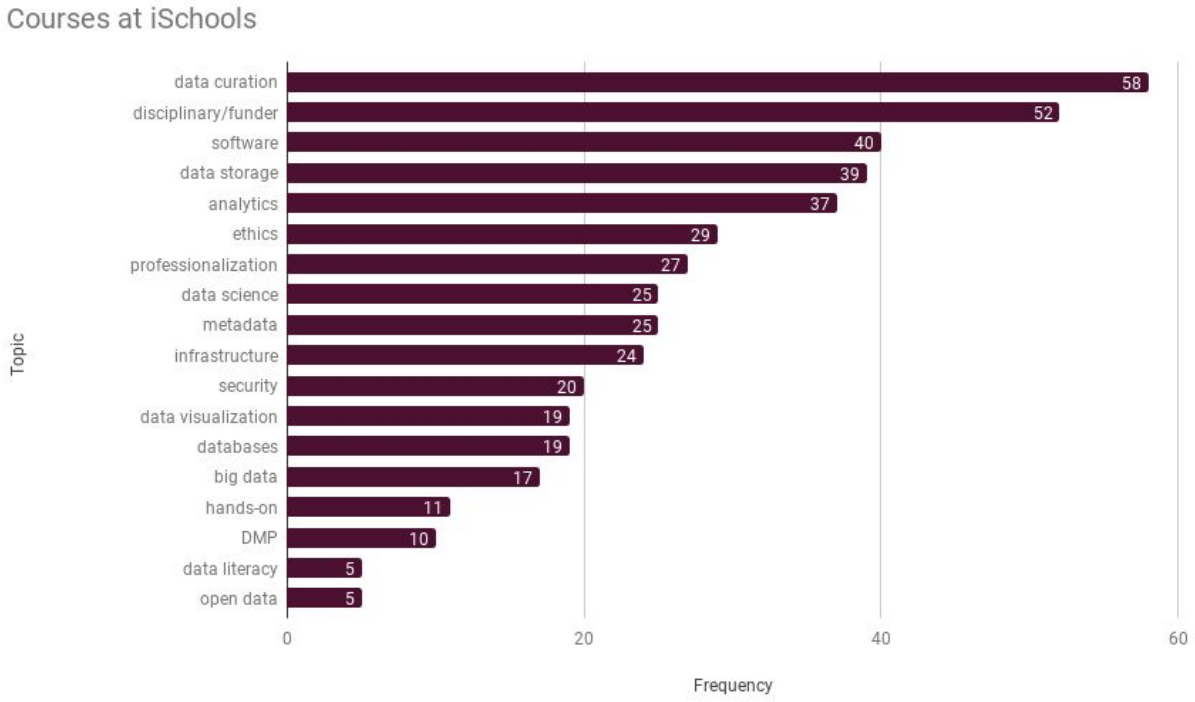


Figure 6. Topics from iSchool courses

Publications

Over 148 publications were gathered in preparation for publishing an annotated research data management bibliography. Of these publications, 82 were articles, 17 were reports or white papers, 14 were papers or posters from conferences, 10 were books, six were policies, five were book chapters, one was an independent study paper, one was a thesis, and one was a podcast episode. Three special issues of journals were identified that were dedicated to research data management in libraries. Seven were bibliographies on research data management were located. Seventeen library guides (i.e., libguides), which can be utilized as a training tool, were identified that covered research data management. Additionally, one data set and one unpublished survey were found that dealt specifically with research data management.

Conclusion

The RDMLA program would be available online with credit. Simmons University would be responsible for instructional design and continuing education credits. As noted from the needs

inventory, the RDMLA courses would be tailorable to multiple proficiency levels and professional requirements. A total of seven modules would be developed:

1. Foundations of Research Data Management (RDM)
2. Navigating research data culture
3. Advocating and marketing the value of RDM in libraries
4. Establishing and managing data services in libraries
5. Overview of coding tools
6. R, Python, Jupyter Notebook
7. Overview of platform tools

Additionally, the Mendeley Data modules developed by Elsevier would be included in the RDMLA.

Next Steps

Within the RDMLA project, the next steps are as follows:

1. RDMLA partners sign up for responsibility in course content area for the seven modules
2. Work with Simmons Online to develop template for online modules
3. Partners develop content using the Simmons Online template
4. Simmons Online uploads the content
5. Work with Simmons University and the Simmons University SLIS program for continuing education (CE) credit opportunities
6. Conduct user-experience (UX) testing of the content and platform
7. Make revisions based on UX testing and feedback
8. Go live by end of 2019

Within the broader project, as the environmental scan via surveys and inventories showed, there are many resources available for RDM training. However, the training is disparate and there is not a universal curriculum. An openly available bibliography could be created from the publications that could benefit the wider research data management community. There is also need for outreach to the librarian community for more input. This outreach could include the datalibs groups, as well as the Research Data Alliance (RDA).

Appendix

Tables

Partners

Boston University Tufts University	Massachusetts College of Pharmacy and Health Science University
Elsevier	Northeastern University
Harvard Library	Simmons University
Harvard Medical School	Tufts University

Table 1. RDMLA partners

Inventory

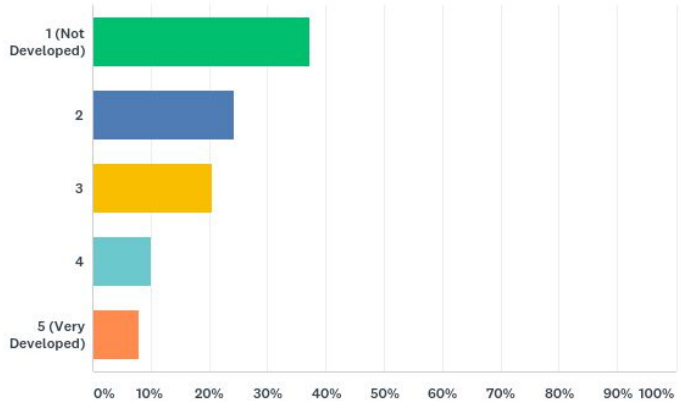
analytics	data science	DMP★	open data
big data	data storage✢	ethics☆	professionalization
consultations/instruction/outreach	data visualization	infrastructure	security✧
data curation*	disciplinary/funder requirements	metadata	software

*Data curation: management throughout data lifecycle
✢Data storage: data repositories, preservation
★DMP: data management plan(s)/planning
☆Ethics: relates to data; includes privacy
✧Security: includes sharing

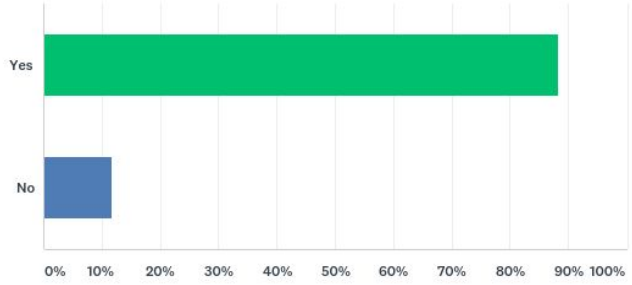
Table 2. Competencies listed

SurveyMonkey Results

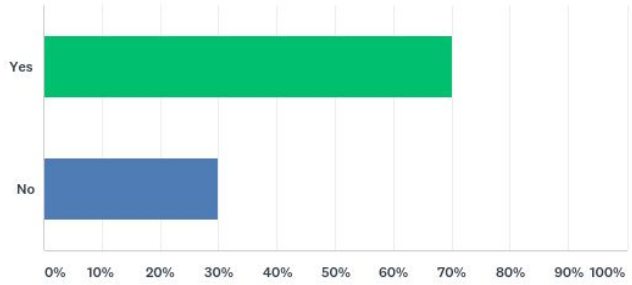
Q1 How developed is your role with your institution's research data management (RDM)?



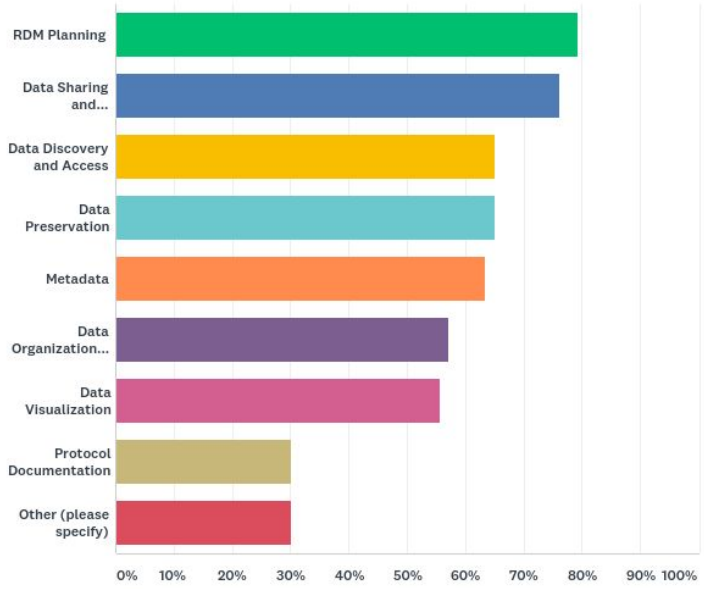
Q3 Do you wish you had a more formal role with RDM?



Q6 Do you personally feel prepared to offer RDM services?



Q16 What RDM services does your institution offer? (Select all that apply)



Q18 What are the primary RDM tools used by your institution?