

Data literacy training needs of biomedical researchers

Lisa M. Federer, MLIS; Ya-Ling Lu, PhD; Douglas J. Joubert, MS

See end of article for authors' affiliations.

DOI: <http://dx.doi.org/10.3163/1536-5050.104.1.008>

Objective: The research investigated topic priorities for data literacy training for biomedical researchers and staff.

Methods: An electronic survey was used to assess researchers' level of knowledge related to data literacy skills and the relevance of these skills to their work.

Results: Most respondents did not have any formal training in data literacy. Respondents considered most tasks highly relevant to their work but rated their expertise in tasks lower.

Conclusion: Among this group, researchers have diverse data literacy training needs. Librarians' expertise makes them well suited to provide such training.

Keywords: Data Curation; Education, Continuing; Library Services; Libraries, Medical; Biomedical Research; Data Collection; Information Storage and Retrieval; Information Dissemination

Many libraries have taken on the role of providing instruction in data literacy, which can be defined as the set of skills and knowledge that “enables individuals to access, interpret, critically assess, manage, handle and ethically use data” [1, 2]. Librarians' expertise and training in skills like metadata, searching and discovery, archiving and preservation, and knowledge management should make them ideal partners for researchers who need to learn to apply these skills to their own data. Libraries' central role to the research process also makes them an ideal place in the research enterprise to house data services and related instruction efforts.


A growing body of literature has addressed a variety of efforts aimed at providing data literacy instruction at biomedical and health sciences libraries [3–7]. Resources like the New England Collaborative Data Management Curriculum provide a helpful set of teaching resources, which librarians can customize to the needs of their audience [5, 8]; however, these approaches and the related literature have some limitations. First, many libraries have directed their training efforts toward

students at the undergraduate and graduate level, rather than focusing on training for postgraduate researchers [3, 5, 7, 9]. In addition, many library-based training efforts focus specifically on data management or writing data management plans (DMPs), which are only part of the skills and knowledge that constitute the broader concept of data literacy [3, 10].

This article reports on an exploratory study that expands on the existing literature by considering the data literacy training needs of career-level researchers and related staff. This study was conducted to inform the development of the data services program at the National Institutes of Health (NIH) Library, which serves staff at NIH and other agencies in the Department of Health and Human Services.

METHODS

First, the authors investigated whether researchers had previously received data literacy training. Second, we identified priorities for data literacy instruction by identifying the skills that researchers considered most relevant to their work, as well as the skills in which they judged their expertise as being lower. Finally, we aimed to determine

 A supplemental appendix and supplemental Table 1, Table 2, Figure 1, and Figure 2 are available with the online version of this journal.

whether differences existed in skill relevance and expertise in different groups in the community based on job role.

Data were collected through a four-section survey consisting of twenty-one questions. The full survey instrument is available in the online only appendix. The survey instrument provided definitions of skills to ensure respondents understood the questions and used terminology common in the scientific research community, rather than library-specific terminology. For example, because researchers would likely be unfamiliar with the concept of “data literacy” training, we used the term “data management,” which was likely more familiar to respondents. The survey instrument was tested in a pilot study and revised accordingly. The NIH Office of Human Subjects Research Protections determined that this survey did not require review by an institutional review board (IRB). In lieu of IRB review, the director, NIH Office of Research Services, approved the final survey instrument.

The first section of the survey contained nine pairs of questions that asked respondents to rate their experience with specific data literacy skills and the relevance of each skill to their work, using a five-point Likert scale, from “Very low” to “Very high.” Each rank was assigned a numerical value for analysis (1=“Very low,” 2=“Low,” 3=“Medium,” 4=“High,” 5=“Very high”). The second and third sections of the survey considered respondents’ attitudes toward and experience with sharing research data. The fourth section contained questions about respondents’ demographics and class scheduling preferences.

The survey was designed to elicit information on two related but substantively different topics: data literacy training needs and data sharing practices. We combined these two topics into one survey to reduce survey burden and maximize the information yielded by the survey [11]. Because the survey results provide insight into two independent areas of inquiry, we report the results of the two sections separately. This article reports on findings related to data literacy training only; findings about data sharing are reported elsewhere [12].

The data literacy training section of the survey considered nine key skills covering a variety of competencies across the research process, from the planning stage to end-of-project tasks like preservation and retention. These skills and the definitions provided in the survey are:

- **Metadata:** Capture and create metadata (descriptive information about your data, how it was collected, and other contextualizing information)
- **Ontology:** Use common data elements, ontologies (formal models of concepts in a domain and their relationships), or other predefined terms for describing your data or variables
- **Collaboration:** Organize, tag, and track data so multiple team members can work on the same dataset
- **Data mining:** Conduct research through data mining (using computational methods to discover patterns in large datasets)
- **Reuse:** Locate and obtain other researchers’ shared data to use in your research, and clean or process it to meet your research needs
- **Visualization:** Demonstrate, analyze, or communicate your research results through data visualization
- **Retention:** Create a plan for long-term storage and retention of your data
- **Deposit:** Publish and deposit data in a repository suited to your research field
- **DMP:** Write a formal DMP, including selecting file formats, choosing a standard for data description, and planning for storage and preservation

Throughout this article, skills are referred to by these shortened names for simplicity.

Respondents were recruited through announcements to various NIH email distribution lists, and the survey was promoted on the NIH Library’s website and digital displays in and near the library. Responses were collected electronically using the NIH Library’s licensed version of SurveyMonkey, an online survey tool. A total of 190 responses were collected during a period of 50 days in April and May 2014. Because the survey was announced in multiple outlets to increase the response rate, the number of potential respondents cannot be estimated. However, the sample is small, representing about 3% of the 6,000 employees in NIH’s Intramural Research Program. Twenty of the respondents did not indicate their position category and were therefore excluded from analyses, leaving a total of 170 eligible responses. All potentially identifiable information was removed before analysis, and both descriptive and inferential statistical analyses were used to examine variables and the potential relationships among these variables. Figures were created with R [13] and RStudio [14], using ggplot2 [15].

RESULTS

Respondent demographics

Respondents were asked to classify themselves according to their primary focus of work or research, selecting the most appropriate response from 3 categories. Thirty-five respondents (21%) identified as “administrative, management, and support staff,” which could include individuals who provided research support for NIH intramural researchers, as well as managers and supervisors. This category would also include NIH staff who administered extramural funding activities. Twenty-two respondents (13%) identified as “clinical research staff,” who worked directly with patients or whose work had clinical applications, such as design of pharmaceuticals or medical devices. One hundred thirteen respondents (67%) identified as “basic science researchers,” whose work focused on preclinical trials, such as in vitro or animal studies, as well as computational research.

Q1. Have researchers previously received relevant training?

The majority of respondents overall (77%), as well as in each position category, responded that they had never had any formal training, with scientific research staff reporting the lowest rates of previous training (Table 1, online only).

Q2. What data literacy skills are priorities for curriculum development?

Ratings for relevance of skills to work and level of expertise in each skill were used to guide curriculum development. The median ranking for relevance and expertise in each skill was calculated; skills with a high median relevance (suggesting a generally high level of interest among the respondents) or a low median expertise (suggesting a generally low level of knowledge among the respondents) are considered a high training priority.

Respondents considered most of the skills highly relevant to their work but rated their expertise in all tasks as medium or lower. Overall, visualization was ranked the most relevant (median=5) and DMP the least relevant (median=3) to respondents’ work. Median expertise was lowest for DMP and ontology (median=2 for both tasks). Figure 1, online only, demonstrates the overall distribution of responses,

and Table 2, online only, displays median relevance and expertise.

Q3. Do relevance and expertise differ by job role?

Rank medians were also calculated for each of the three position category subgroups to determine whether instruction priorities differed based on position category. Table 3 displays median relevance and expertise rankings for each position category. Figure 2, online only, contains the distribution of responses divided by position category.

DISCUSSION

The high proportion of respondents who indicated that they had never had formal data literacy training demonstrates a need for training opportunities. Our results also indicate that respondents find a variety of data literacy skills relevant to their work but do not necessarily have a correspondingly high level of expertise. The finding that median expertise for all tasks was medium or lower, both overall and within each position category subgroup, suggests a need for training that addresses each of the skills considered in this study. Visualization, which had the highest overall median *relevance* ranking, and ontology and DMP, which had the lowest overall median *expertise* ranking, can be considered priorities for instruction for this audience.

For most skills, fewer than one-fifth of respondents ranked their expertise as “Very low,” indicating that most respondents had at least some knowledge of these skills and did not need a course intended for complete beginners. However, the broad range of ratings of expertise in a given task also suggests that multiple levels of instruction may best meet the needs of researchers with differing skill levels. With such an approach, clear descriptions of learning outcomes and class topics would help ensure that researchers are able to decide which class is most appropriate for their level of expertise.

Our study also suggests that researchers working in different areas of the research enterprise are not completely homogenous in their data literacy training needs, as differences in median relevance and expertise exist across the three position category subgroups for several of the skills. Some of these differences could be explained by these subgroups’ different work roles and the types of data that they utilize. For example, clinical researchers frequently

	Scientific		Clinical		Administrative	
	Median relevance	Median expertise	Median relevance	Median expertise	Median relevance	Median expertise
Metadata	4	3	4	3	4	3
Ontology	4	2	4	2.5	3.5	2
Collaboration	4	3	4	2	4	2
Data mining	4	3	4	2	4	2
Reuse	4	3	3	2	4	2
Visualization	5	3	4	3	4	3
Retention	4	3	4	3	4	2
Deposit	4	3	3	2	3.5	3
Data management plan	3	2	3	2	3	2

1="Very low," 2="Low," 3="Medium," 4="High," 5="Very high".

Table 3

Median relevance and expertise rankings by position category

work with patient data that may contain personally identifiable information and therefore are prohibited from freely sharing their data, which could explain this subgroup's lower ranking for deposit expertise and reuse relevance (Table 3) [12]. Other differences cannot be readily explained by our data or by dissimilarities inherent to the three subgroups, such as scientific researchers rating visualization as more highly relevant than their clinical and administrative colleagues. Future research could be helpful in validating our findings and elucidating the reasons for differences if they persist in larger studies.

Pilot testing specialized training sessions designed for specific segments of the research community might also be reasonable. These training sessions could feature the topics that are most relevant to that group and draw on case studies, examples, and exercises similar to what attendees are likely to encounter in their daily work. Skills rated as less relevant could be viewed as a low training priority, but lower relevance ratings could also suggest a need to communicate the importance of these skills to researchers. For example, writing a DMP was rated as the least relevant skill, with respondents overall as well as each subgroup ranking it medium in relevance, likely because NIH does not currently require researchers to prepare a DMP. However, NIH's response to the 2013 Office of Science and Technology Policy's memo on access to federally funded research indicates that "NIH is taking steps to ensure all NIH-funded researchers develop data management plans whether they are funded by a grant, cooperative agreement, contract, or intramural funds, regardless of funding level," and that all relevant policies will be enacted by the end of calendar year 2015 [16, 17]. Thus, among NIH-

funded researchers, the perception that DMPs are of somewhat low relevance to their work may change within the next year. Librarians and other information professionals may want to proactively address the increased importance of DMPs. Other policy and technology changes can cause researchers' perception of a skill's relevance to shift over time. Therefore, librarians who provide data literacy training might find it helpful to remain up to date with such developments in order to provide timely and relevant classes.

Data literacy skills are relevant to researchers' daily work and are well within the scope of librarians' expertise. To support researchers' changing needs in the face of a rapidly evolving research enterprise that increasingly relies on data literacy skills, libraries may want to consider providing training suitable not only for students, but also for career-level researchers and staff. A single workshop is clearly no substitute for the years of training and experience that librarians have in skills like metadata and ontologies, preservation, and information management, so libraries could also investigate services other than training to assist researchers in meeting policy requirements and improving data management. Providing library-based training programs and data services for researchers will likely increase their awareness of librarians' skills and create new opportunities for librarians to engage with this population.

Limitations

Our sample size was small and based on convenience sampling, limiting the generalizability

of our results. Given that recruitment was conducted primarily through email lists for the NIH Library and data-related groups at NIH, selection may be biased in favor of individuals who already have an interest in data literacy and who may consider these skills more relevant to their work than individuals who do not have such interests. The NIH research community may not be representative of the population of biomedical researchers on the whole, since researchers who choose to work in government research settings may differ from their peers who work in private or academic settings. Finally, this study relies on respondents' self-assessment of their expertise in tasks, because of the difficulties in quantitatively measuring data literacy knowledge in a brief online survey. Respondents' rating of their own expertise might not correlate with their actual expertise, as people frequently overestimate their abilities in self-assessments [18].

Further research is needed to investigate whether the findings of this study are applicable to the broader research community. Studies that assess the needs of researchers at different career stages, as well as in different areas of specialization, can be helpful in enabling librarians and others engaged in teaching data literacy to create customized and effective curricula. As researchers are increasingly expected to produce datasets that are well managed, clear, understandable, and shareable with the scientific community, targeted training based on established needs can play an important role in ensuring researchers' success.

REFERENCES

1. Calzado Prado J, Marzal MÁ. Incorporating data literacy into information literacy programs: core competencies and contents. *Libri*. 2013 Jun;63(2):123–34. DOI: <http://dx.doi.org/10.1515/libri-2013-0010>.
2. Koltay T. Data literacy: in search of a name and identity. *J Documentation*. 2015;71(2):401–15. DOI: <http://dx.doi.org/10.1108/jd-02-2014-0026>.
3. Adamick J, Reznik-Zellen R, Sheridan M. Data management training for graduate students at a large research university. *J eScience Lib*. 2012;1(3):180–8. DOI: <http://dx.doi.org/10.7191/jeslib.2012.1022>.
4. Carlson JR, Fosmire M, Miller C, Sapp Nelson M. Determining data information literacy needs: a study of students and research faculty libraries faculty and staff scholarship and research. 2011:paper 23. (preprint).
5. Eaker C. Planning data management education initiatives: process, feedback, and future directions. *J eScience Lib*. 2014. DOI: <http://dx.doi.org/10.7191/jeslib.2014.1054>.
6. Darabi H, McCue K, Beesley J, Michailidou K, Nord S, Kar S, Humphreys K, Thompson D, Ghousaini M, Bolla MK, Dennis J, Wang Q, Canisius S, Scott CG, Apicella C, Hopper JL, Southey MC, Stone J, Broeks A, Schmidt MK, Scott RJ, Lophatananon A, Muir K, Beckmann MW, Ekici AB, Fasching PA, Heusinger K, Dos-Santos-Silva I, Peto J, Tomlinson I, Sawyer EJ, Burwinkel B, Marme F, Guénel P, Truong T, Bojesen SE, Flyger H, Benitez J, González-Neira A, Anton-Culver H, Neuhausen SL, Arndt V, Brenner H, Engel C, Meindl A, Schmutzler RK; German Consortium of Hereditary Breast and Ovarian Cancer, Arnold N, Brauch H, Hamann U, Chang-Claude J, Khan S, Nevanlinna H, Ito H, Matsuo K, Bogdanova NV, Dörk T, Lindblom A, Margolin S; kConFab/AOCS Investigators, Kosma VM, Mannermaa A, Tseng CC, Wu AH, Floris G, Lambrechts D, Rudolph A, Peterlongo P, Radice P, Couch FJ, Vachon C, Giles GG, McLean C, Milne RL, Dugué PA, Haiman CA, Maskarinec G, Woolcott C, Henderson BE, Goldberg MS, Simard J, Teo SH, Mariapun S, Helland Å, Haakensen V, Zheng W, Beeghly-Fadiel A, Tamimi R, Jukkola-Vuorinen A, Winqvist R, Andrulis IL, Knight JA, Devilee P, Tollenaar RA, Figueroa J, García-Closas M, Czene K, Hooning MJ, Tilanus-Linthorst M, Li J, Gao YT, Shu XO, Cox A, Cross SS, Luben R, Khaw KT, Choi JY, Kang D, Hartman M, Lim WY, Kabisch M, Torres D, Jakubowska A, Lubinski J, McKay J, Sangrajrang S, Toland AE, Yannoukakos D, Shen CY, Yu JC, Ziogas A, Schoemaker MJ, Swerdlow A, Borresen-Dale AL, Kristensen V, French JD, Edwards SL, Dunning AM, Easton DF, Hall P, Chenevix-Trench G. Polymorphisms in a putative enhancer at the 10q21.2 breast cancer risk locus regulate NRBF2 expression. *Am J Human Genetics*. 2015 Jul 2;97(1):22–34. DOI: <http://dx.doi.org/10.1016/j.ajhg.2015.05.002>.
7. Piorun M, Kafel D, Leger-Hornby T, Najafi S, Martin E, Colombo P, LaPelle NR. Teaching research data management: an undergraduate/graduate curriculum. *J eScience Lib*. 2012;1(1):46–50. DOI: <http://dx.doi.org/10.7191/jeslib.2012.1003>.
8. University of Massachusetts Medical School Lamar Soutter Library. New England collaborative data management curriculum [Internet]. *The Library*; 2011 [cited 15 Mar 2015]. <<http://library.umassmed.edu/necdmc>>.
9. Dekker H. Using web-based software to promote data literacy in a large enrollment undergraduate course. Presented at: World Library and Information Congress: 76th IFLA General Conference and Assembly; Gothenburg, Sweden; 2010.
10. Johnston L, Lafferty M, Petsan B. Training researchers on data management: a scalable, cross-disciplinary approach. *J eScience Lib*. 2012;1(2):article 2. DOI: <http://dx.doi.org/10.7191/jeslib.2012.1012>.

11. Olson CA. Survey burden, response rates, and the tragedy of the commons. *J Continuing Educ Health Professions*. 2014 Spring;34(2):93–5. Epub 2014/06/19. DOI: <http://dx.doi.org/10.1002/chp.21238>. PubMed PMID: 24939350.
12. Federer LM, Lu YL, Joubert DJ, Welsh J, Brandys B. Biomedical data sharing and reuse: attitudes and practices of clinical and scientific research staff. *PLOS ONE*. 2015. DOI: <http://dx.doi.org/10.1371/journal.pone.0129506>.
13. R Core Team. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2014.
14. RStudio. *RStudio: integrated development environment for R (version 0.98.994)*. Boston, MA: RStudio; 2012.
15. Wickham H. *ggplot2: elegant graphics for data analysis*. New York, NY: Springer; 2009.
16. National Institutes of Health. NIH public access plan for increasing access to scientific publications and digital scientific data from NIH funded scientific research [Internet]. The Institutes; Feb 2015 [cited 16 Mar 2015]. <<http://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf>>.
17. Office of Science and Technology Policy. Increasing access to the results of federally funded scientific research [Internet]. The Office; 2013 [cited 12 Aug 2014]. <http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf>.
18. Dunning D, Heath C, Suls JM. Flawed self-assessment: implications for health, education and the workplace. *Psychological Sci Public Interest*. 2004 Dec;5(3):69–106. DOI: <http://dx.doi.org/10.1111/j.1529-1006.2004.00018.x>.

AUTHORS' AFFILIATIONS



Lisa M. Federer, MLIS, lisa.federer@nih.gov, Research Data Informationist; **Ya-Ling Lu, PhD**, ya-ling.lu@nih.gov, Informationist; **Douglas J. Joubert, MS**, joubertd@ors.od.nih.gov, Informationist; NIH Library, Division of Library Services, Office of Research Services, National Institutes of Health, 10 Center

Drive, Bethesda, MD, 20892

Received June 2015; accepted August 2015