RESEARCH REPORT

Starting the data conversation: informing data services at an academic health sciences library

Kevin B. Read, MLIS, MAS; Alisa Surkis, PhD, MLS; Catherine Larson, MS; Aileen McCrillis, MS, MPH, AHIP; Alice Graff; Joey Nicholson, MLIS, MPH; Juanchan Xu, MD, MS

See end of article for authors' affiliations.

DOI: http://dx.doi.org/10.3163/1536-5050.103.3.005

Objective: The research obtained information to plan data-related products and services.

Methods: Biomedical researchers in an academic medical center were selected using purposive sampling and interviewed using open-ended questions based on a literature review. Interviews were conducted until saturation was achieved.

Results: Interview responses informed library planners about researchers' key data issues.

Conclusions: This approach proved valuable for planning data management products and services and raising library visibility among clients in the research data realm.

Keywords: Libraries, Medical; Biomedical Research; Data Collection; Information Storage and Retrieval; Health Information Management; non-MeSH: Data management; medical libraries; health sciences libraries; interview methods

INTRODUCTION

Over the past five to ten years, libraries have begun to provide data-related services to researchers. Examples include assisting researchers in complying with the data management and sharing requirements of federally funded grants (e.g., National Institutes of Health, National Science Foundation) [1–4]; providing guidance for developing workflows and standard data collection procedures [1–4]; training researchers on how to better organize, store, and preserve their data [1, 3, 4]; and building searchable interfaces to provide a level of discovery and access for research datasets [1, 3–5]. Health sciences libraries, however, have been slow to develop these services [6, 7]. New data sharing and data management initiatives from the National Institutes of Health's Big Data to Knowledge initiative and the publishers Public Library of Science (PLOS) [8, 9] have created new opportunities for librarians, in particular for health sciences librarians, to expand their roles in data services.

A number of libraries have assessed researchers' data-related issues and needs as a way to guide the development of their services in this area. These assessments have been done through interviews [10–16], focus groups [17–21], and web audits or bibliographic analyses [22–29]. However, few have addressed these data needs in the context of health sciences research [18, 20, 23] or provided a methodology that the authors found satisfactory for gathering information about researchers' data management practices.

This paper describes the methodology that the authors used to identify researchers to interview, reach out to those researchers, and conduct the interviews. It describes key findings from the interviews about the challenges that researchers face when collecting and managing data.

METHODS

The authors, located in an academic health sciences library, completed a series of interviews as a means to assess their research community and the challenges that the researchers face when collecting, managing, storing, and preserving their research data. These interviews were also designed to build connections with the researcher community. They were intended to provide valuable information to plan the development of library products and services, including an institutional data catalog to describe researchers' datasets created at the medical center, and led to the development of a tool to help basic science labs better manage their research data.

Developing interview questions

We performed a literature review to identify studies that evaluated the data-related challenges and needs of an institution's researchers. The library then selected a number of interview questions from previous studies that were deemed most appropriate for understanding researchers' data management challenges [12, 16, 18, 20, 26]. Questions taken from previous studies were adapted to make the interviews more conversational and open-ended. Additional interview questions were developed by the library to create a conversational interaction (Appendix A online only). The rationale was that if the interviews had a conversational tone, researchers would be more likely to elaborate on their answers, providing more in-depth information and bringing to light issues

This article has been approved for the Medical Library Association's Independent Reading Program <http://www.mlanet.org/education/irp/>.

Supplemental Appendix A, Appendix B, and Appendix C are available with the online version of this journal.

about which the librarians would not have thought to ask, due to the differences between their perspective and the researchers'.

Selecting study participants

Researchers with active grant funding were selected. Data from the institution's grants management tool were used to identify eligible participants. The grants management tool retrieved data from institutional researchers including their administrative department, grant funding agency, grant title, and contact information. Using the data gathered from the grants management tool, the authors identified and purposively selected researchers based on their expected data service needs, types of research (e.g., basic science, clinical research), levels of research experience, and involvement in big data research. Selected participants were sent an email outlining the librarians' intention to learn more about their data-related needs. Two attempts were made to reach out to researchers, after which a lack of response resulted in the researcher's removal from the list of potential interviewees. The authors interviewed individual researchers until theoretical saturation was achieved, such that no new insights into key requirements for library data services were identified.

Conducting the interviews

Prior to each interview, the librarians reviewed the stated research interests and publications of the researchers being interviewed to gain a better understanding of their research methods, including the types of data collected, the data collection methods used, and whether the researchers used newly created data or existing data from previous studies. This information provided librarians with the necessary background to feel confident discussing researchers' data during the interviews and provided context for the interviewer as the researchers responded to questions about their research data.

Two librarians were present for each interview: one who led the discussion and another who took notes on a laptop using word processing software. Using two librarians allowed the interview to remain conversational, so that one librarian would not be tasked with asking questions, listening intently, and taking notes at the same time.

Analyzing the results

Notes collected during the interviews were saved to a secure institutional server, and no personal identifying information was collected; only the distinction between basic science and clinical researchers was recorded, as well as the researchers' departments. Interview responses were coded in a word processing document using the grounded theory method and then transferred to a spreadsheet with an indication of being collected from either a basic science or clinical researcher (Appendix B online only). This spreadsheet served as a large, de-identified dataset, comprising frequencies of the major themes related to the data management of the interviewed researchers. The institutional review board gave this study an exemption, as the de-identified dataset categorized this study as non-human subject research.

RESULTS

Researchers were invited to participate in the study until theoretical saturation was achieved, at which time the authors had conducted thirty interviews, comprising eleven interviews with basic scientists and nineteen with clinical researchers. A number of responses to the questions were unique to individual researchers and therefore did not provide the librarians with information they could use to implement widespread products and services. These results can be viewed in the online Appendix C. Themes that did emerge from the interviews are described in Table 1. The specific themes that provided the library with an opportunity to implement new products and services are discussed in more detail below.

Data organization challenges and needs

Basic science researchers. The basic science researchers interviewed identified several challenges in managing their data. The biggest obstacle for researchers was the perceived lack of standards and procedures available for them to uniformly collect their data. Without specific collection standards, researchers were left to develop custom data collection methods, constantly reinventing the wheel, sometimes with every new research project.

Another issue that researchers identified was a disconnect between the different types of data collected. For example, imaging data and raw numerical data that were collected as part of the same research project were often located in different places and, therefore, difficult to find. Postdoctoral researchers and graduate students, who work in a lab for a limited amount of time, exacerbate this problem: these researchers work on a specific project but then leave with either the physical data or the methodology they used to collect that data. This leaves the basic science researcher without the ability to understand *who* used their data, *how* they used their data, or *where* their data have gone once that researcher leaves.

Clinical researchers. The major challenges identified by clinical researchers related to the quality of their data. Many researchers mentioned data quality as a major concern. This issue often stems from the involvement of multiple personnel in collecting data for a clinical study, coupled with inconsistent data collection methods. These inconsistencies can result in team members entering data elements using different interpretations of a given variable (e.g., weight measured in pounds versus kilograms), potentially

Table 1
Results from data interviews

	Basic scientists (n=11)	%	Clinical researchers (n=l9)	%	Overall (n=30)	%
Data storage methods						
Data repository	2	(18%)	1	(5%)	3	(10%)
Institutional server	5	(45%)	18	(95%)	23	(77%)
External hard drive	5	(45%)	5	(26%)	10	(33%)
DVD	2	(18%)	1	(5%)	3	(10%)
Drop box	3	(27%)	3	(16%)	6	(20%)
External institute	1	(9%)	4	(21%)	5	(17%)
Others with single responses (Appendix B)						
File formats used						
Lab notebook/paper	7	(64%)	7	(37%)	14	(47%)
Excel	5	(45%)	9	(47%)	14	(47%)
Comma separated values	1	(9%)	1	(5%)	2	(7%)
Others with single responses (Appendix B)						
Data organization methods						
Documented procedures	1	(9%)	5	(26%)	6	(20%)
Data dictionary	—	—	7	(37%)	7	(23%)
Folders	4	(36%)	5	(26%)	9	(30%)
Paper cheat sheet	—	—	1	(5%)	1	(3%)
Lab notebook	3	(27%)	—	—	3	(10%)
Shared drive	2	(18%)	—	—	2	(7%)
Willingness to reuse data (their own and other people's research data)						
Yes	3	(27%)	15	(79%)	18	(60%)
No	2	(18%)	1	(5%)	3	(10%)
For comparison only	4	(36%)	—	—	4	(13%)
Only their own data for use in future studies	6	(55%)	15	(79%)	21	(70%)
Challenges of data organization						
Poor data output formats	—	—	5	(26%)	5	(17%)
Data quality	—	—	4	(21%)	4	(13%)
Disparate datasets	5	(45%)	2	(11%)	7	(23%)
Team miscommunication	—	—	2	(11%)	2	(7%)
Lack of standards	7	(64%)	—	—	7	(23%)
Postdoc/student leaves with data	5	(45%)	—	—	5	(17%)
Too time consuming	5	(45%)	—	—	5	(17%)
Cannot search data	—	—	1	(5%)	1	(3%)
Data loss	—	—	1	(5%)	1	(3%)
Size of data	1	(9%)	—	—	1	(3%)
Interest in data sharing						
Sharing with the public	3	(27%)	11	(58%)	14	(47%)
Sharing via collaboration only	5	(45%)	6	(32%)	11	(37%)
Not interested in sharing	3	(27%)	2	(11%)	5	(17%)

rendering a data element or an entire dataset useless. Clinical researchers also identified difficulties in transferring data from one format to another. Clinical researchers use a number of different types of statistical software (e.g., SAS, SPSS, STATA, R) as part of their research process, and moving data between different types of software often results in poor data quality and even data loss.

Researcher interest in data sharing

Identifying researchers most interested in sharing their data was essential to inform the implementation of a data catalog for internally generated research datasets. The interviews identified clinical researchers—particularly those in the Department of Population Health (11 researchers)—as willing to share their data with the public as long as they were aware of who was using their data. Those same researchers expressed interest in finding shared datasets for their own research, either through direct access or collaboration. Responses to the interviews suggested that basic science researchers currently show little interest in sharing their research data, as the majority preferred to share with their direct collaborators or with no one at all. Basic science researchers cited a number of reasons for a reluctance to share data including negative experiences with past sharing, concerns about privacy restrictions, the belief that their data are too specialized to be of value to others, insufficient storage options for sharing data publically, and the hurdle of having to organize their data prior to sharing.

DISCUSSION

The biggest challenge that libraries face in building data management services is the researchers' perception that librarians do not understand research data and have no role to play in data management. While several other studies interviewed researchers about their data management challenges, many took an approach that seemed to call upon the researchers to be conversant in the language of the library, rather than speaking to the researchers in their own language. For example, the use of terminology such as ''e-science,'' ''metadata,'' and ''Dublin Core'' throughout the data interview process—terms that have little to no meaning for most researchers—may serve to widen, rather than narrow, the gap between librarians and researchers. Through the careful construction of ''researcher-centric'' questions and thorough preparation by the interviewers in educating themselves about the researchers' work, the interviewers were able to avoid this potential pitfall.

Another strategy the librarians found to be very effective was making the interviews conversational and open-ended. Providing a relaxed environment for the researchers allowed the interview questions to flow more coherently, gave the librarians the opportunity to ask the researchers to elaborate on their answers in a more natural way, and allowed room for the researchers to expand their answers into areas that the librarians, with their different perspective, might have overlooked.

Through the data interviews, the authors gained valuable knowledge about the medical center research community's data issues including, but not limited to, the challenges they face when collecting, organizing, and sharing their research. Insights gained from the interviews provided new information that led to the improvement or development of library data products and services. The understanding that the Department of Population Health is most keen to share their data and find other research datasets that they can use for their research provided useful information that allowed the library to build out its data catalog to first address the needs of its most likely users. The data interview results regarding the extent of the difficulties that basic science researchers face in organizing the data in their labs led to the development of a low-barrier lab organization tool that is currently being piloted in two basic science labs.

Data interviews are an effective means of elucidating the challenges that researchers at an institution face when collecting, organizing, and sharing their data. The interviews also raise the visibility and, when conducted well, can enhance the credibility of the library in the realm of research data. Because of both benefits of raised visibility and credibility and the high variability of responses across researchers and so presumably across institutions, the value of what is reported in this report may lie more in the methodology than the specific results, as these interviews can serve as an important first step for a health sciences library to insert itself into the data conversation and change the perceptions of the research communities that they support.

REFERENCES

1. Borgman CL, Wallis JC, Enyedy N. Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. Int J Digit Lib [Internet]. 2007;7:17–30 [cited 12 Feb 2015]. <http://escholarship.org/uc/item/6fs4559s#>.

2. Carlson JR. Demystifying the data interview: developing a foundation for reference librarians to talk with researchers about their data. Lib Res [Internet]. 2011 [cited 12 Feb 2015]. <http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1186&context=lib_research>.

3. Gold A. Cyberinfrastructure, data, and libraries, part 1: a cyberinfrastructure primer for librarians. D-Lib Mag [Internet]. 2007 Sep/Oct;13(9/10) [cited 12 Feb 2015]. <http://www.dlib.org/dlib/september07/gold/09gold-pt1.html>.

4. Gold A. Cyberinfrastructure, data, and libraries, part 2: libraries and the data challenge: roles and actions for libraries. D-Lib Mag [Internet]. 2007 Sep/Oct;13(9/10) [cited 12 Feb 2015]. <http://www.dlib.org/dlib/september07/gold/09gold-pt2.html>.

5. Hey T, Trefethen A. The data deluge: an e-science perspective. In: Berman F, Fox G, Hey AJG, eds. Grid computing: making the global infrastructure a reality [Internet]. Chichester, UK: Wiley; 2003. p. 1–17 [cited 12 Feb 2015]. <http://eprints.soton.ac.uk/257648/1/The_Data_Deluge.pdf>.

6. Creamer A, Morales M, Crespo J, Kafel D, Martin ER. Data curation and management competencies of New England Region health sciences and science and technology librarians [Internet]. Presented at: University of Massachusetts and New England Area Librarian e-Science Symposium 2011 [cited 12 Feb 2015]. <http://escholarship.umassmed.edu/escience_symposium/2011/posters/8>.

7. Creamer A, Morales M, Crespo J, Kafel D, Martin ER. Assessment of health sciences and science and technology librarian e-science educational needs to develop an e-science web portal for librarians. J Med Lib Assoc. 2011 Apr; 99(2):153–6. DOI: http://dx.doi.org/10.3163/1536-5050.99.2.007.

8. PLOS One. Editorial and publishing policies: sharing of data, materials, and software [Internet]. PLOS; 2014 [cited 5 Jan 2015]. <http://www.plosone.org/static/policies#sharing>.

9. Nature Communications. Availability of data, materials and methods [Internet]. Nature Publishing Group; 18 Nov 2014 [cited 3 Jan 2015]. <http://www.nature.com/authors/policies/availability.html>.

10. Carlson J, Fosmire M, Miller CC, Nelson MS. Determining data information literacy needs: a study of students and research faculty. portal Lib Acad. 2011; 11(2):629–57.

11. Jones S, Ross S, Ruusalepp R. Data audit framework methodology [Internet]. Glasgow, UK: 2009. p. 1–70 [cited 12 Feb 2015]. <http://www.data-audit.eu/DAF_Methodology.pdf>.

12. Lage K, Losoff B, Maness J. Receptivity to library involvement in scientific data curation: a case study at the University of Colorado Boulder. portal Lib Acad [Internet]. 2011 Oct;11(4):915–37 [cited 21 Nov]. <http://muse.jhu.edu/journals/portal_libraries_and_the_academy/v011/11.4.lage.html>.

13. Peters C, Dryden AR. Assessing the academic library's role in campus-wide research data management: a first step at the University of Houston. Sci Technol Lib [Internet]. Routledge; 2011 Sep;30(4):387–403. DOI: http://dx.doi.org/10.1080/0194262X.2011.626340.

14. Raboin R, Reznik-Zellen RC, Salo D. Forging new service paths: institutional approaches to providing research data management services. J eScience Lib [Internet]. 2012;1(3) [cited 12 Feb 2015]. <http://escholarship.umassmed.edu/jeslib/vol1/iss3/2/>.

15. Walters TO. Data curation program development in U.S. universities: the Georgia Institute of Technology example.

Int J Digit Curation [Internet]. 2009;4(3):83–92 [cited 12 Feb 2015]. <http://www.ijdc.net/index.php/ijdc/article/viewFile/136/153>.

16. Westra B. Data services for the sciences: a needs assessment. Ariadne [Internet]. 2010;(64) [cited 12 Feb 2015]. <http://www.ariadne.ac.uk/issue64/westra>.

17. Adamick J, Canavan M, McGinty S, Reznik-Zellen R, Schmidt M, Stevens R. Building as we climb: the data working group at the University of Massachusetts Amherst [Internet]. Presented at: University of Massachusetts and New England Area Librarian e-Science Symposium 2011 [cited 12 Feb 2015]. <http://escholarship.umassmed.edu/escience_symposium/2011/posters/3>.

18. Bardyn TP, Resnick T, Camina SK. Translational researchers' perceptions of data management practices and data curation needs: findings from a focus group in an academic health sciences library. J Web Lib [Internet]. 2012 Oct;6(4):274–87 [cited 30 Jan 2013]. <http://www.tandfonline.com/doi/abs/10.1080/19322909.2012.730375>.

19. Delserone LM. At the watershed: preparing for research data management and stewardship at the University of Minnesota Libraries. Lib Trends. 2008 Fall57(2): 202–10.

20. Johnson LM, Butler JT, Johnston LR. Developing e-science and research services and support at the University of Minnesota Health Sciences Libraries. J Lib Adm [Internet]. Routledge; 2012 Nov;52(8):754–69. DOI: http://dx.doi.org/10.1080/01930826.2012.751291.

21. Trinidad SB, Fullerton SM, Bares JM, Jarvik GP, Larson EB, Burke W. Genomic research and wide data sharing: views of prospective participants. Genetics Med Off J Am Coll Med Genetics. 2010 Aug; 12(8):486–95.

22. Harrison A, Searle S. Not drowning, ingesting: dealing with the research data deluge at an institutional level. VALA2010 Proceedings [Internet]. 2010 [cited 12 Feb 2015]. <http://www.vala.org.au/vala2010/papers2010/VALA2010_43_Harrison_Final.pdf>.

23. Hruby GW, McKiernan J, Bakken S, Weng C. A centralized research data repository enhances retrospective outcomes research capacity: a case report. J Am Med Inform Assoc. 2013 Jan 15;1–5. DOI: http://dx.doi.org/10.1136/amiajnl-2012-001302.

24. Newton MP, Miller CC, Bracke MS. Librarian roles in institutional repository data set collecting: outcomes of a research library task force. Collect Manag. 2011; 36(1):53–67.

25. Reznik-Zellen R, Adamick J, McGinty S. Tiers of research data support services. J eScience Lib [Internet]. 2012;1(1):27–35 [cited 10 Nov 2012]. <http://escholarship.umassmed.edu/jeslib/vol1/iss1/5/>.

26. Scaramozzino JM, Ramirez ML, McGaughey KJ. A study of faculty data curation behaviors and attitudes at a teaching-centered university. Coll Res Lib. 2012 Jul 1; 73(4):349–65.

27. Soehner C, Steeves C, Ward J. E-science and data support services: a study of ARL member institutions [Internet]. Association of Research Libraries; 2010 [cited 11 Jan 2013]. <http://www.arl.org/storage/documents/publications/escience-report-2010.pdf>.

28. Williams SC. Using a bibliographic study to identify faculty candidates for data services. Sci Technol Lib [Internet]. 2013 May 9;32(2);202–9. DOI: http://dx.doi.org/10.1080/0194262X.2013.774622.

29. Xia J, Liu Y. Usage patterns of open genomic data. Coll Res Lib [Internet]. 2013 Mar 1;74(2):195–207 [cited 7 Mar 2013]. <http://crl.acrl.org/content/74/2/195.abstract>.

AUTHORS' AFFILIATIONS

Kevin B. Read, MLIS, MAS (Principal Investigator), kevin.read@nyumc.org, Knowledge Management Librarian; **Alisa Surkis, PhD, MLS,** alisa.surkis@nyumc.org, Translational Science Librarian; **Catherine Larson, MS,** Catherine.Larson@med.nyu.edu, Web Services Librarian; **Aileen McCrillis, MS, MPH, AHIP,** Aileen.mccrillis@med.nyu.edu, Research Librarian/User Experience Librarian; **Alice Graff**, Alice.Graff@nyumc.org; **Joey Nicholson,** MLIS, MPH, Joey.nicholson@med.nyu.edu, Education and Curriculum Librarian; **Juanchan Xu, MD, MS,** Juanchan.xu@med.nyu.edu, Ontology Manager; Health Sciences Libraries, New York University, 577 First Avenue, New York, NY 10016