JESLIB 2012; 1(1): 18-26
doi:10.7191/jeslib.2012.1006

Journal of eScience Librarianship
putting the pieces together: theory and practice

An Assessment of Needed Competencies to Promote the Data Curation and Data Management Librarianship of Health Sciences and Science and Technology Librarians in New England

Andrew Creamer, Myrna E. Morales, Javier Crespo, Donna Kafel, Elaine R. Martin

University of Massachusetts Medical School, Worcester, MA, USA

Abstract

Purpose: The purpose of this study was to evaluate health sciences and science and technology librarians' needed data curation and management (DCM) competencies to support nascent and future patron and institutional eScience research endeavors. The data from this research will be used to align a data curation and management curriculum with the educational needs of an online eScience portal community of users, and create relevant future professional development for librarians interested in data curation and eScience librarianship.

Setting/Participants: The study targeted the needed data curation and data management competencies of health sciences and science and technology librarians in six U.S. states who are on a listserv of librarians interested in learning about eScience. The sample for this study was 63 librarians.

Methodology: The team created the assessment tool using content analyses of digital curation and management library literature and LIS data management curricula. The survey contained 15 open-ended and closed-ended questions and was distributed to 141 librarians using SurveyMonkey (http://www.surveymonkey.com).

Results/Outcomes: The team identified twenty needed competency areas related to data curation and data management. The participants identified the necessary competencies to provide data curation and data management services. Results revealed a small number of librarians engaged in DCM and infrequent data services requests. Findings suggest there is an increase in libraries pursuing strategic plans concerning data management and the library community needs to cultivate a diverse range of technical and non-technical competencies through future professional development. Librarians saw their future roles involving DCM and sought competencies in conducting data interviews with patrons and helping patrons with NSF data management requirements. The survey results indicate the greatest need for librarians is technical hands-on training in the digital description and curation of large data sets.

Discussion/Conclusion: Librarians are interested in developing data curation and data management competencies to support eScience. These data indicate that future relevant professional development for librarians interested in eScience should focus on non-technical and technical DCM competencies.

Correspondence to Andrew Creamer: acreamer@vt.edu
Keywords: Data management, Data curation, eResearch, Competencies, eScience Librarianship

Introduction

The Association of Research Libraries *Agenda for Developing E-Science in Research Libraries* (2007) characterized eScience as "well-established experimental and theoretical methodologies with large-scale, data-driven, and computationally intense characteristics. E-science fundamentally alters the ways in which scientists carry out their work, the tools they use, the types of problems they address, and the nature of the documentation and publication that results from their research. E-science requires new strategies for research support and significant development of infrastructure" (ARL 2007, 6). Integral to eScience, data curation and management (DCM) involves providing or helping provide storage and access to these large data sets accumulated by researchers.

DCM is a relatively new field for health science librarians (ARL 2007; Rambo 2009). Since it strives to provide ways for data to be easily available to other researchers, the field is very Internet dependent. This may require traditional library skills such as cataloging and description, developing search procedures for large data sets, and exploring ways to merge sets in a meaningful way. Anna Gold (2010) advocated for establishing the "legitimacy of library roles in data curation through formal education and training as well as by integrating data curation into existing library services" (Gold 2010, 23).

In 2011 the University of Massachusetts Medical School Lamar Soutter Library and National Network of Libraries of Medicine New England Region started an online eScience educational portal (http://esciencelibrary.umassmed.edu) specifically for supporting librarians' eScience and DCM roles (Martin and Kafel 2010; Creamer, Morales et al. 2011). The portal has benefited from many pioneering librarians' published experiences founding eScience and digital curation and management initiatives within their institutions (Angevaare 2009; Arms, Calimlim et al. 2009; Choudhury 2008; Garritano and Carlson 2009; Gold 2007a; Gold 2007b).

Yet among this body of research, the portal team found little data regarding the needed competencies for data curation and management. This included detailed information about the nature and frequency of data services that librarians are providing and the data-related competencies they possess or will need to provide these services. One of the aims of the portal is to teach DCM skills to librarians who are not currently providing these services so that that will be able to do so in the future. Thus, the portal team realized if it were to accurately align the portal's data management curriculum with its community's needed competencies, it would need to identify which DCM competencies the community felt it needed in order to cultivate DCM skills, establish DCM roles, and engage successfully in eScience. This paper reports the results of that survey and outlines the DCM competencies currently being sought by this sample of health sciences and science and technology librarians in New England.

Methods

The team developed the DCM competencies survey questions using content analyses of data services-related terms within digital library position postings, eScience, and data curation and data management literature between 2007 and 2011. It also analyzed selected Library and Information Science programs' data curation and management curricula. The team used the Digital Curation Centre's (2010) Data Curation Lifecycle Model as the themes for analysis. The survey contained fifteen questions, nine closed-ended and six open-ended questions. After receiving IRB exemption, the team distributed the survey using SurveyMonkey. A small group of librarians at the University of Massachusetts Medical School tested the survey and offered feedback to the team. The team then sent the revised survey in the spring of 2011 to 124 New England health sciences

and science and technology librarians on its eScience Listserv and 17 Library Directors via the Boston Library Consortium's Listserv. After six weeks, the team collected and analyzed the results. The sample for this study was sixty-three librarians (See Electronic Content on page 8).

Results

Sixty-three of the 141 librarians (44.68%) responded. Twenty librarians (31.7%) worked at a health sciences library, 39 (61.9%) at a non-health sciences affiliated academic library, and 4 (6.3%) at a special library. Forty-three librarians (67.2%) had more than 10 years of experience. Thirteen librarians (20.6%) were currently providing digital data management and curation services, and 21 (33.3%) stated they would be in the future. Thirty librarians (47.6%) did not currently provide DCM services, 43 (69.4%) stated their library has developed or is in the process of creating a data management plan for the library or related library services policies or strategic plans. The majority stated the plans were in the early stages. Most plans called for establishing an eScience position or data management group to provide data services to patrons.

When asked to list competencies and skills used to curate and manage data sets or those they would need to provide these services, 28 librarians (62.2%) listed the use of Web 2.0 technologies (Table 1). Eighteen librarians (40.0%) listed the design and maintenance of digital databases, and 13 (28.9%) listed the use of programming and scripting languages. Eighteen librarians (40.0) listed the maintenance of institutional data repositories, and 19 (42.2%) listed the use of specific metadata standards. Thirteen librarians (28.9%) listed the manipulation of these standards, creation of crosswalks and data description, indexing and storage. Seventeen librarians (37.8%) listed the provision of data mining, interpretation, representation and visualization services, 20 (44.4%) listed the provision of data archiving and preservation services, 10 (22.2%) listed the development of digital lab notebook applications, and 19 (42.2%) listed the development of virtual tools to manage and curate data. Nine librarians (20.0%) did not provide any of these data services.

The team asked respondents to list the data services they perform or would like to perform for patrons (Table 2). Thirty-two librarians (72.7%) listed conducting a "data interview" with researchers to assess their data needs. Twenty-five librarians (56.8%) listed working with researchers to create a data management plan, and 32 (72.7%) listed consulting with researchers on the life cycle of their data. Twenty-six librarians (59.1%) listed teaching data literacy to patrons and 30 (68.2%) listed helping researchers be compliant with grant data management requirements. Thirty-one librarians (70.5%) listed helping patrons understand the intellectual property issues concerning data. Ten librarians (22.7%) listed working with researchers on data security issues and 32 (72.7%) listed advertising their data services. Thirty-five librarians (79.5%) listed the promotion of data sharing and open access, and encouraging researcher participation in the institutional repository. Twenty-five librarians (56.8%) listed accessing data sets from published literature for patrons' original research. One librarian (2.3%) listed none of the above.

The team asked respondents to suggest additional competencies for the portal's curriculum. One volunteered understanding the taxonomy of data management and curation. Another offered an awareness of types of data and their usage. One suggested the ability to identify and build collaborations with researchers, the Institutional Review Board (IRB), the Clinical and Translational Science Award (CTSA), and Information Technology (IT) groups. Respondents suggested competencies in policy issues involving intellectual property and privacy. Another recommended digital project management skills: "these skills have been essential for

my work, especially cleaning metadata and publishing digital collections." Other suggestions were cyberinfrastructure competencies, using applications across platforms, and domain knowledge and the principle research problems specific to those domains.

The team asked respondents to list the data management and curation skills that they would like to gain or improve through continuing education. The largest request was foundational knowledge. Specifically, respondents wanted hands-on data management and curation instruction. Others requested advanced instruction addressing discipline-specific metadata, metadata for data, and instruction on linking data using the semantic web, and demonstrations of the tools to mine these data. In addition, they requested instruction on the various ways researchers use data sets and transition active data to archived data. Others requested ways to gauge the data needs of patrons and the access to more open-source data management tools. Many highlighted the need for continuing education addressing intellectual property issues concerning data.

Among the respondents who described the frequency of the patron requests for data services, the responses ranged from "none" and "few" to as frequently as "monthly" and "biweekly." One respondent described advertising a workshop on data management and, at the time of the survey, already had 56 patrons registered. A few, who responded that such requests were still infrequent, described feeling that their library had not successfully promoted its services, and consequently many patrons and faculty were still unaware of such services. The majority of these requests concerned grant-related data management requirements. For example, one librarian stated that shortly after his or her institution had established a data management group, it quickly received researcher requests for assistance with their NSF data management plans. Yet this same respondent noted that data services requests from non-researcher patrons were "very low, perhaps once or twice per year." Similarly, the other respondents noted that the NSF requirement increased the number of patron data services requests.

The survey asked respondents to comment on the obstacles slowing the establishment of eScience services. Respondents lamented the lack of funds to train staff, or hire or attract new staff with data management skills. They described a lack of funds to upgrade or purchase cyberinfrastructure resources. Many felt that librarians, already overwhelmed, were finding it difficult to keep up with the speed of evolving technology, policy, and data management and curation continuing education. Others noted that a lack of institutional policies regarding data management created confusion or overlap regarding the responsibilities of data curation. Others stated that their library management was not prioritizing data management or there was a lack of institutional support, while others noted a lack of patron awareness and lack of researchers' trust in the library's ability to maintain and secure their data.

Discussion

The survey findings provided much information on the current state of DCM and eScience in the region's libraries. The first major finding is that only a small percentage of librarians are actually engaged in the digital curation and management of large data sets and patron data services requests are infrequent. This finding can be viewed from multiple perspectives. Some might feel that this is low in light of the region's high number of Category I research universities, while others might feel this is high considering the field of eScience librarianship is still in its infancy.

The second major finding is that more than half of the respondents' libraries were actively engaged in creating a library strategic plan or policy for data management. The team expected this result in light of the high num-

Table 1: Technical data management and curation competencies organized from greatest to lowest response

Technical Competencies	Response Percent	Response Count
I use Web 2.0 technologies	62.2%	28
I provide data archiving and preservation services	44.4%	20
I work with and/or develop virtual tools to manage and curate data	42.2%	19
I work with a variety of metadata standards (e.g. interoperability standards and language such as Dublin Core, MODS, and OAI-PMH, etc.)	42.2%	19
I build, populate and maintain digital databases	40.0%	18
I maintain an institutional repository	40.0%	18
I provide data mining, interpretation, representation and visualization services	37.8%	17
I work with metadata manipulation, crosswalk, validation and portals (e.g. description, indexing, storing, etc.)	28.9%	13
I use a variety of programming languages (e.g. XML, SQL, etc.)	28.9%	13
I work with and/or develop digital lab notebook applications	22.2%	10

ber of participants interested in grant-related data plans such as those prescribed by the NSF. These data management plans offer an opportunity for librarians to establish an additional role to support research within their institutions.

The third major finding is that currently librarians lack the technical skills needed to manage and curate terabytes of digital data. This result confirms the need for continuing education and library school curricula to emphasize hands-on curation and management of large scientific data sets, and also confirms the need for online tools like the portal to help practicing librarians gain these skills. These low DCM technical skill results provide valuable targets for the portal curriculum.

The fourth major finding was the high interest in cultivating competencies regarding the conducting of data interviews and expertise in data literacy and intellectual property issues. These results were expected in light of the NSF data plan requirements and the increasing need for librarians to educate faculty and researchers on these requirements

Table 2: Non-technical data management and curation competencies organized from greatest to lowest response

Non-Technical Data Competencies	Response Percent	Response Count
I promote digital data sharing, open access, and/or participation in an institutional repository at my institution	79.5%	35
I actively advertise my and the library's data services to researchers at my institution	72.7%	32
I perform a "data interview" with researchers to assess their data needs at various stages of their research	72.7%	32
I consult with researchers about the life cycle of their data and work with them on archival and conservancy issues prior, during and post project	72.7%	32
I help patrons understand the intellectual property and copyright issues concerning their data (e.g. provenance, publication, licensing and digital rights)	70.5%	31
I work with researchers to help them be compliant with government-sponsored grants' regulations and requests concerning data management (e.g. NSF)	68.2%	30
I teach data literacy to patrons	59.1%	26
I work with researchers to create a data management plan before they begin data collection/aggregation	56.8%	25
I access or locate data sets from the published literature for patrons' original research papers	56.8%	25
I work with researchers on data security issues	27.7%	10

and the ways that the library can help them with regulatory compliance and understanding their digital rights.

The final major finding was the serious barriers facing librarians and libraries trying to engage in eScience. While the team expected financial constraints for hiring and purchasing equipment, it did not expect the high number of responses regarding institutional barriers such as lack of support and the territorial struggles between IT and various other institutional departments. The team hopes to use these findings to create curricular competencies and goals to increase DCM and eScience librarianship. Although one of the limits to this study is that these data represent health sciences and science and technology librarians in six U.S. states and cannot infer national or global concerns, the portal is online and available for use by any librarian, anywhere in the world, who is interested in eScience librarianship.

Conclusion

Health sciences and science and technology librarians in New England are very interested in providing DCM services and would like to cultivate a number of the competencies outlined above to provide these services; the majority of respondents stated their library has or is in the process of creating a data management policy and this could be driving their interest in learning about DCM and developing competencies to support these initiatives. Indeed, this would support recent research showing that grant-related data management requirements will increase librarian interest and engagement in e-science (Hswe and Holt 2011). Participants are acutely aware of the competencies they will need to successfully serve research patrons: data literacy and technical competencies. In order to construct a data management plan, librarians will require data literacy competencies concerning data lifecycle and preservation, intellectual property and scholarly communication issues related to data, and researchers' data requirements. In order to manage and curate these data sets, they will need to cultivate cyberinfrastructure and technical competencies to build and manage a data repository, manipulate metadata, and ensure system interoperability (Johnston 2010).

The data curation and management competencies outlined here are goals for professional development. One does not need to possess all before engaging in eScience. On the contrary, as one survey respondent prudently remarked, "I think there are a wide range of skills needed and not one person is going to have them all; I think it's really a team effort." This sentiment echoes the words of T. Scott Plutchak (2011). In his 2011 Janet Doe lecture, he exhibited a slide of a recently advertised data services library position. The posting had an ambitious list of sought-after skills for just one librarian, causing him and many in attendance to chuckle. Remarking on this ambition, he turned to the audience and his smiling face turned into one of serious concern:

> "I fear it represents a sense that we cannot stop doing all of the things that our librarians are currently doing in order to address the challenges of dealing with digital materials, so we are going to create one position and get some smart and energetic librarian who can handle everything associated with digital. And then the rest of us can continue doing the essential jobs that we are doing and not have to worry about all that weird stuff. But if we are not all thinking of ourselves as digital services librarians, we are in trouble" (Plutchak 2011, 16).

The website for the e-Science Portal for New England Librarians officially went live in 2011. The survey's participants have established a framework of needed skills that will be used to guide the portal's future data curation and management competencies, however, they have also provided the profession with insight on the progress of its data curation and management initiatives. This assessment has shown that starting these initiatives come with challenges. In addition to financial and staffing challenges, librarians are struggling with establishing an institutional role in data management and deploying the technical infrastructure necessary for storage and preservation of immense data sets. Therefore, the portal team plans to investigate how librarians are managing and curating data within these constraints. The University of Massachusetts Medical School Lamar Soutter Library and National Network of Libraries of Medicine New England Region are using the survey data to organize future librarian professional development programming including symposia and professional days focused on data management and curation. It also hopes to assess how the portal is assisting health sciences librarians in their efforts to integrate data management and curation into their practice and embrace an inevitably digital future.

Electronic Content

Appendix: Survey Instrument
An online supplement to this article can be found at http://escholarship.umassmed.edu/jeslib/vol1/iss1/4/ under "Additional Files".

References

Angevaare, Inge. 2009. "Taking care of digital collections and data: "curation" and organisational choices for research libraries." *Liber Quarterly: The Journal of European Research Libraries* 19: Accessed March 15, 2011, http://liber.library.uu.nl/publish/articles/000278/article.pdf.

Arms, William , Manuel Calimlim, and Lucia Walle. 2009. "Escience in practice: lessons from the Cornell web lab." *D-Lib Magazine*, 15: Accessed March 15, 2011, http://www.dlib.org/dlib/may09/arms/05arms.html.

Association of Research Libraries, "Joint Task Force on Library Support for E-Science. Agenda for developing e-science in research libraries," *Association of Research Libraries* (2007). Accessed January15, 2010, http://www.arl.org/bm~doc/ARL_EScience_final.pdf.

Choudhury, Sayeed. 2008. "Case study in data curation at Johns Hopkins University." *Library Trends* 57: 211-20. Accessed March 16, 2011, https://jscholarship.library.jhu.edu/handle/1774.2/34023.

Creamer, Andrew, Myrna Morales, Javier Crespo, Donna Kafel, and Elaine Martin. 2011. "Assessment of health sciences and science and technology librarian e-science educational needs to develop an e-science web portal for librarians." *Journal of the Medical Library Association* 99: 153-5.

Digital Curation Centre, accessed April 1, 2011, http://www.dcc.ac.uk/digital-curation/what-digital-curation.

Garritano, Jeremy, and Jake Carlson. 2009. "A subject librarian's guide to collaborating on e-science projects." *Issues in Science & Technology Librarianship* 57: Accessed March 16, 2011, http://www.istl.org/09-spring/refereed2.html.

Gold, Anna. 2007. "Cyberinfrastructure, data, and libraries, part 1 a cyberinfrastructure primer for librarians." *D-Lib Magazine*. Accessed March 16, 2011, http://www.dlib.org/dlib/september07/gold/09gold-pt1.html.

Gold, Anna. 2007. "Cyberinfrastructure, data, and libraries, part 2: libraries and the data challenge: roles and actions for libraries." *D-Lib Magazine*. Accessed March 16, 2011, http://www.dlib.org/dlib/september07/gold/09gold-pt2.html.

Gold, Anna. 2010. "Data curation and libraries: short-term developments, long-term prospects." Accessed March 15, 2011, http://works.bepress.com/agold01/9.

Hswe, Patricia, and Ann Holt. 2011. "Joining in the enterprise of response in the wake of the NSF data management planning requirement." *Research Library Issues* 274: 11-17.

Johnston, Lisa. 2010. "User-needs assessment of the research cyberinfrastructure for the 21st century." *IATUL Proceedings* vii: 63-78.

Martin, Elaine R, and Donna Kafel. 2010. "Response to Neil Rambo's editorial: "E-science and biomedical libraries." *Journal of the Medical Library Association* 98: 5.

Plutchak, T. Scott. "Breaking the barriers of time and space: the dawning of the great age of librarians," Medical Library Association Janet Doe Lecture Medical Library Annual Meeting Minneapolis, MN, accessed May 16, 2011, http://www.mlanet.org/awards/honors/doe.html.

Rambo, Neil. 2009. "E-science and biomedi-

cal libraries." *Journal of the Medical Library Association* 97: 159–61.

Acknowledgement

The authors would like to thank the editors and journal's reviewers for their insightful feedback and the members of the E-science Listserv, and the board members, staff and online user community of the E-science Portal for New England Librarians.

Funding Statement

This project has been funded by the National Library of Medicine, National Institutes of Health, Department of Health and Human Services, under contract no. N01-LM-6-3508 with the University of Massachusetts Medical School.

Disclosure: The authors report no conflicts of interest.

ISSN 2161-3974