December 14, 2012

Data Curation is for Everyone! The Case for Master's and Baccalaureate Institutional Engagement with Data Curation

Yasmeen Shorish, *James Madison University*

Data Curation is for Everyone!

The Case for Master's and Baccalaureate Institutional Engagement with Data Curation

Yasmeen Shorish[1]

[Keywords]data curation, data management, libraries, institutional repository, higher education

[Abstract]This article describes the fundamental challenges to data curation, how these challenges may be compounded for smaller institutions, and how data management is an essential and manageable component of data curation. Data curation is often discussed within the confines of large, research universities. As a result, master's and baccalaureate institutions may be left with the impression that they cannot engage with data curation. However, by proactively engaging with faculty, libraries of all sizes can build closer relationships and help educate faculty on data documentation and organization best practices. Experiences from one master's comprehensive institution as it engages with data management can provide guidance on how to begin the conversation and plan for future engagement with data curation. In a period of several months, James Madison University went from no formal data advising to a coordinated data management support plan for faculty. Collaboration across campus—and across institutions—can help make data curation an accomplishable goal. Comparisons to scalable efforts with institutional repositories are made to further encourage participation with data curation. Funding agency mandates are not the only cause for engagement with research data. Research universities account for 297 of the 1832 four-year institutes of higher education in the United States of America. There must be stewardship of the data from the remaining 1535 organizations to help preserve the complete intellectual product of the nation. Additional resources and readings are included.

Introduction

Libraries provide access to information and increasingly, that information is found in the form of digital data sets. Many research universities (classified as RU by the Carnegie Classification) have responded to this production of digital data with concerted efforts to manage these data. The challenge to preserve and make accessible the vast quantities of digital data being produced is one that has caused librarians—and their institutions—some angst as they try to determine the best course of action. It is not uncommon for large, research libraries to have working groups or full-time positions dedicated to areas such as scholarly communication, research data, or copyright issues. However, master's and

[1] Science Librarian, James Madison University MSC 4601 Harrisonburg, VA 22801. shorisyl@jmu.edu

baccalaureate institutions that lack a research mandate may find themselves outside of the digital data discussion. Smaller academic libraries may not have the personnel available to staff those departments, yet their faculty still need support in those areas. Identifying that need and explaining to administration that a topic such as data curation should be addressed by the library is a challenge to many organizations where the librarians are already performing several job duties. Libraries should be the custodians of digital data, just as they have served as stewards of the record in the print realm. Librarians possess the skill set to document, organize, and make data discoverable. Research data is increasingly called on to be made more open and sharable (National Science Foundation 2010, VI-8) and librarians can help meet that demand.

Compounding this challenge is the difficult nature of digital data itself. Digital data exists in many formats, each with its own idiosyncrasies. Data can consist of image files, text files, proprietary sensor logs, or computer code. Its permutations are limitless. Additionally, data requires maintenance to ensure its usability. Checksum operations, back-ups, and migrations (both software and hardware) are necessary procedures in the preservation of these materials, especially if one wishes to access and use the data in the future. In addition to these issues, the technological landscape is in a state of perpetual, rapid evolution. Taking these factors together, one can begin to see why many libraries may hesitate to address the issue of digital data management and curation. However, Anna Gold has argued that while the curation of digital data complicates the libraries' mission of stewardship, it has an immense reward: "a well-curated record of scientific data could itself become a new, vital and useful part of the process and practices of science, whether through data-mining and reuse, data-visualization, or other techniques and methods yet to be invented" (2010, 5). Bryan Heidorn (2011) further reinforces this view, arguing that it is the role of libraries to collect and distribute the intellectual gains of society and preserve the scientific record of data that cannot be easily replicated.

Background

Curation often means different things to different audiences and there is active discussion around this ambiguity in the profession. Sarah Shreeves and Melissa Cragin define data curation as "the active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education, which includes appraisal and selection, representation and organization of these data for access and use over time" (2008, 90). Wrapped up in that statement are several complex concepts: active management, lifecycle, and usefulness. How active should "active management" be? What

constitutes a lifecycle, and who makes that decision? How does one determine usefulness? These questions, familiar to archivists, can be answered with archival theory and the application of selection and appraisal. Ross Harvey identifies three key differences in the appraisal and selection decisions between digital data and analog materials in the "Appraisal and Selection" chapter of the *Digital Curation Centre Digital Curation Manual*. These key areas are:

- the technical ability to preserve data;
- the ongoing cost of data maintenance; and
- the need to make preservation decisions early on in the lifecycle of the data, before the data becomes unusable or inaccessible (2006, 10).

The first two areas can present the most immediate challenge to smaller institutions that lack resources, be they financial or personnel. However, the third area is one with which any institution can be actively engaged. While preservation decisions that are made will be affected by the institution's ability to cover costs of maintenance and technology, all institutions should be encouraging researchers to think about data preservation as close to the point of data creation as possible. Educating researchers about the benefits of data management is one method of encouragement and can serve as a stepping stone into the field of data curation.

Data Management: a first step

Data management involves planning for the data prior to data collection, discussing security, and determining sharing responsibilities. The planning stage can be an area of heavy library involvement. While initially time consuming, developing a data management planning procedure leads to conversations with faculty and streamlines the data organization process as faculty engage with the library's efforts. There are many factors to be considered in data management, such as any funding requirements, ethical or legal concerns, and the documentation and organization of the data. Researchers can often identify their responsibilities when it comes to funding requirements, such as National Science Foundation (NSF) or National Institutes of Health (NIH) mandates, or ethical concerns, such as identifiable health data. Thinking about the kinds of data that will be produced, how and when it will be produced, and where it will be kept once it is produced should be considered the first steps in managing the data. However, Michael Witt, et al. (2009) found that many research faculty were unsure of how best to organize and manage their data in order to make it understandable and usable by others.

This is why librarian expertise in employing descriptive standards and the stewardship of information is a vital asset in the management of data.

Working with faculty, librarians – especially subject specialists – can provide guidance on this issue. Consistent file naming protocols, secure storage (with back-ups), and a plan for preserving and accessing the data are all ways that a researcher can ensure that his or her data is findable and replicable in the future. Good data management can be protective for a researcher whose findings are called into question or for a researcher whose work utilizes student labor, which can change year to year. Research conducted with sound data management practices can help ensure that all members of the research team are 'speaking the same language' throughout the data collection and analysis period. Moreover, data management can help ease the transition from the storage of research data to the curation of that data, once the organization is prepared to engage at that level.

Why Curate

As previously mentioned, many large, research institutions are actively engaged with data curation due to the volume of research data generated on campus. Such institutions have established working groups and departments to advise researchers on data management and to determine research data preservation needs. Examples include Cornell University's Research Data Management Support Group, Johns Hopkins University's Data Management Group, and Purdue University's Distributed Data Curation Center. These institutions, among others, have invested time, money, and personnel to provide high levels of data services to their communities and to examine the shifting landscape of data management, preservation, and curation.

The question is do smaller, teaching-centric institutions need to be concerned with data curation? Should librarians at these institutions be involved? The answer to these interconnected questions is ideally, yes. While the quantities of research data at smaller institutions may be less, the quality of the data and its usefulness to others is still relevant. In order to preserve that relevant data for potential future use, some level of curation must be applied to it. Research conducted at smaller, non-RU institutions often falls into the category of "small science." Small science refers to hypothesis-driven research, often with limited funding, led by a principal investigator and a small team, usually students (Cragin et al. 2010, 4024). This research may be part of a research network, like the Howard Hughes Medical Institute's Science Education Alliance, or it could be confined within a department. This level of

research can have immediately identifiable impacts, such as a research network working towards a national or global concern; or the impacts could slowly become evident, such as when a department's research has curriculum development implications. While master's and baccalaureate institutions may not have dedicated working groups from the library available to address data preservation and access concerns, there are opportunities for some level of library engagement with these researchers and issues. Tracy Gabridge (2009) clearly details the possible avenues of participation for liaison librarians and curation efforts, including the identification of data repositories, data management planning, teaching data literacy, data set collection building, and the creation of data preservation standards. Engagement in some or all of these areas can provide the faculty with valuable support and guidance, especially when trying to meet any funding requirements.

Small Scale Efforts: a work in progress

At James Madison University (JMU), a master's comprehensive, public institution in Virginia, there was an identified need for data management guidance due to the NSF's data management plan mandate. In 2011, the NSF issued a requirement that all grant applications contain a two page data management plan (DMP). JMU is a teaching institution with a student population of over 19,000, primarily undergraduates. There has always been a strong focus on undergraduate research and JMU has been a NSF Research Experiences for Undergraduates site for several years. Moreover, the past ten years has seen an increase of over 60% in grant awards at the federal, state, and private levels (James Madison University Office of Sponsored Programs 2011, 1). However, some faculty were unsure how to approach the NSF DMP requirement. After investigating faculty concerns, and with the arrival of a new science librarian in the fall of 2011, the library began to formally engage with the issue of data management.

Data management represents a small part of the data curation landscape. However, the present situation at JMU is one where a formalized curation workflow cannot yet be implemented. There is no institutional repository or scholarly communication librarian and the complexities of data curation extend beyond the purview of one liaison librarian. It was for these reasons that engaging with data management, with the understanding of its implications for the larger issue of data curation, was deemed the most feasible and sustainable undertaking for JMU. Through conversations with faculty, guided in part through Purdue University's Data Curation Profiles Toolkit (http://datacurationprofiles.org/), JMU began forming an understanding of researcher needs and

planning for the directions that the university may need to move in the future. The science librarian asked physics and biology faculty several questions, including: what kinds of data they were producing, did they anticipate it being useful to others outside of JMU, for what period of time did they think it would be useful, and how would they like to see it stored (locally vs. disciplinary repository). As each research project is unique, each researcher had a different opinion on every question, although they were most comfortable with keeping the data on campus and mediating its sharing.

Providing faculty with links to other universities' data curation websites felt like a half step, so the library organized a workshop for faculty prior to the 2012 spring semester. The primary goal of the workshop was to explain the NSF DMP requirements to faculty and to identify data management opportunities in faculty research that could be more fully supported. By inviting campus information technology (IT) and the Office of Sponsored Programs, it was possible to facilitate a discussion around how data is generated, used, and maintained on campus. While some faculty were unclear on the scope of IT services, it became apparent that a cooperative initiative between the library and IT would provide the highest level of faculty support. While unable to consider the active curation of data at this time, working with IT has identified the possibilities and limits to current data storage and preservation options on campus. Additionally, the Assistant Vice President for Information Technology and the Provost of the university were cognizant of the growing quantity of research data on campus and welcomed the opportunity to support and collaborate with the library on this issue.

After providing some basic education on data management, JMU Libraries had to evaluate storage options, given the absence of an institutional repository. After lengthy discussion with faculty, the science librarian and IT developed accurate language reflecting IT archiving capabilities and researcher responsibilities. It was determined that using a single tool to funnel faculty through to develop a DMP would result in the most consistent wording and provide a concentrated area of support. In some RU institutions, this is achieved through consultation with a data services librarian or a data curation task force. At JMU, the best course of action was to adopt the DMPTool (https://dmp.cdlib.org/), a web portal that helps researchers create DMPs and meet funding requirements through a guided and structured form. As a participating institution, JMU was able to insert stock language regarding the data storage and preservation strategies on campus. This language can be modified over time to reflect the evolving capabilities of the library, and the university, with regard to data curation.

JMU Libraries has an interest in growing its data service to incorporate more curatorial practices, but recognizes that long-term planning must occur in order to ensure that future directions are sustainable. Possible considerations include additional library staff, an institutional repository, a distributed repository in cooperation with other peer institutions, or a researcher targeted service, such as DuraCloud's anticipated "Direct-to-Researcher" service. While research is expected to grow on campus, it is of paramount importance that JMU Libraries provide a sustainable level of support, in collaboration with campus partners. Having laid a strong foundation of data management support, it is the goal of the library to be able to smoothly transition to more data curation strategies as they become implementable.

Scalability

Not all of the strategies employed at JMU will be transferable to other institutions. In order to consider issues of storage, data back-up, and security, information technology and other campus partners must be involved. For some, the infrastructure for data back-up and security may not exist as a campus-wide standard. In those cases, it will be necessary to scale down those actions to an individual level, that is, inform the principal investigator of security best practices, such as storing research data on a separate hard drive and backing up to another stable device. Researchers could also investigate cloud storage as a possibility, after ensuring that security considerations—such as stable back-ups and restricted access—are met. Issues related to data sharing must be discussed with the researchers and any other stakeholders. While many smaller institutions do not have institutional repositories for research data, discipline-specific data repositories do exist, and librarians can guide researchers to these resources for depositing data. Databib (http://databib.org), an annotated bibliography of repositories, was created, in part, to meet this need. Adjusting the level of library engagement and service to the needs of the faculty and the resources of the institution is a reasonable approach for every institution.

This concept of scaling the level of engagement with an issue that has typically been considered RU domain is not a new one. Institutional repositories have made the transition from research institutions to master's and baccalaureate institutions, albeit on a smaller scale. A 2007 study by the MIRACLE (Making Institutional Repositories A Collaborative Learning Environment) Project found that among the master's and baccalaureate institutions surveyed, 53 percent were not engaged with institutional repository planning, 20 percent were in the planning stages, 16 percent were piloting, and 11 percent had implemented institutional repositories (Markey et al. 2008, 160). While the majority of respondents

were not involved with planning for a repository, 47 percent *were involved* at some level. In the same study, undergraduates were identified as the main contributors to the institutional repositories of these smaller institutions. This should not be surprising, as undergraduates are the predominant student population at these institutions and the faculty's primary responsibility is on teaching, not publishing. While the contributing populations at RU and master's and baccalaureate institutions are different, they are serving the needs and goals of their communities. The institutional repositories have been scaled, in both physical size (Nykanen 2011, 10) and scope, to fit the expectations and needs of the institution. There is no reason why this same approach cannot be applied to data curation.

Moving Forward

At institutions currently not engaged with data curation, the first step is to identify if there is a need. Any organization receiving funding from the National Science Foundation has an explicit, mandated need to at least begin work with data management, but there may be less obvious, unidentified needs among researchers. Anyone working on a distributed project across campuses would benefit from a data management plan. Communicating with departments and faculty to assess needs and desires is something that librarians do in liaison work such as collection development. In many ways, this communication process is the same. Reaching out to determine what data is being generated and whether it should be curated requires a cooperative audience and time, but no additional infrastructure or financial investment. Utilizing tools such as the Data Curation Profiles Toolkit and the DMPTool can provide guidance to those for whom the concept of data curation has been an unknown.

If possible, conduct a needs assessment of the faculty. This can be achieved through an online survey querying faculty on their research data production and sharing practices, or librarians could conduct a sampling of faculty interviews to determine if any patterns of need arise. This potentially time intensive and arduous process can help institutions identify the necessary level of support, as well as research trends on campus. Based on these findings or faculty requests, an institution can go in one of many directions, depending on the community requirements and the available resources. The library could decide to add data sets to its institutional repository, if one exists, and make preservation decisions based on that workflow. Alternatively, the library could merely advise researchers on data documentation best practices through an FAQ and direct them to campus server space or disciplinary repositories for storage. This hands-off approach may not achieve a high level of curation to the data, but it does direct researchers to tools available, outside of the library, to organize their research and

make it accessible. Looking to RU institutions that have developed robust data curation services, smaller organizations can adapt that methodology for their communities, most likely scaling down the active curation of the assets due to personnel or technology limitations.

The Importance of Curation

When one takes into account all the considerations of selection and maintenance and the changing landscape of technology, the thought of tackling data curation at a master's or baccalaureate institution can be intimidating. However, data is an important component of the future of scholarly communication (Davis and Vickery 2007, 26) and librarians have a responsibility to provide stewardship of this intellectual product. That responsibility does not diminish simply because an institution is not a research university. While most of the research in this area has come from RU institutions, master's and baccalaureate institutions can engage with data curation and share those experiences with one another to help build a knowledge base to serve the community. This community of practice can produce new tools and strategies that best suit its environment.

To put this discussion in perspective, according to the Carnegie Foundation for the Advancement of Teaching, 297 institutions have been classified as research universities and 1535 have been classified as master's or baccalaureate institutions (2011). Those 1535 institutions may individually contribute small amounts of research data when compared to the research universities, but taken as a whole, the data being produced can be considered a substantial contribution to this nation's intellectual capital. Those 1535 institutions can engage with data curation on some level, however minimal, to ensure that the research data of teaching institutions are not lost or hidden. Working collaboratively, these institutions can ensure that all academic libraries are engaged in good stewardship of digital data. More libraries sharing their experiences will help move the conversation from "we are too small to do that" to "we are investigating our needs and moving forward." That is a small, but crucial step towards accepting the stewardship of an institution's intellectual achievements, in all its various forms.

Additional Reading and Resources

The following resources provide additional information on the role of libraries and data curation, to help inform and inspire those just getting started in the field. While many of the publications involve case studies or experiences at research institutions, the strategies employed there can help smaller colleges and universities identify starting points, or be scaled down to an appropriate workflow.

Brandt, D. Scott. 2011. "Disambiguating the Role of Data Lifecycle Gatekeeper." Paper presented at the Workshop on Research Data Management at Princeton University, Princeton, NJ, July 18-20. www.columbia.edu/~rb2568/rdlm/Brandt_Purdue_RDLM2011.pdf.

Carlson, Jake. 2012. "Demystifying the Data Interview: Developing a Foundation for Reference Librarians to Talk with Researchers about their Data." *Reference Services Review* 40 (1): 7-23. doi:10.1108/00907321211203603.

Choudury, Sayeed. April 2010. "Data Curation." *College & Research Libraries News* 71 (4): 194-196. http://crln.acrl.org/content/71/4/194.full

DataONE. "Best Practices", accessed May 11, 2012, https://www.dataone.org/best-practices.

Goldstein, Sarah and Sarah K. Oelker. 2011. "Planning for Data Curation in the Small Liberal Arts College Environment." *Sci-Tech News* 65 (3): 5-11. http://jdc.jefferson.edu/scitechnews/vol65/iss3/4.

Heidorn, P. B. 2011. "The Emerging Role of Libraries in Data Curation and E-Science." *Journal of Library Administration* 51 (7): 662-672. doi:10.1080/01930826.2011.601269.

Heidorn, P. Bryan. 2008. "Shedding Light on the Dark Data in the Long Tail of Science." *Library Trends* 57 (2): 280-299. doi:10.1353/lib.0.0036

Lamar Soutter Library, University of Massachusetts Medical School and George C. Gordon Library, Worcester Polytechnic Institute. 2012. *Frameworks for a Data Management Curriculum*. http://library.umassmed.edu/data_management_frameworks.pdf.

Macdonald, Stuart and Luis Martinez-Uribe. 2010. "Collaboration to Data Curation: Harnessing Institutional Expertise." *New Review of Academic Librarianship* 16: 4-16. doi:10.1080/13614533.2010.505823.

Ogburn, Joyce L. 2010. "The Imperative for Data Curation." *Portal: Libraries and the Academy* 10 (2): 241-246. doi: 10.1353/pla.0.0100.

Provost's Task Force on the Stewardship of Digital Research Data. 2012. *Research Data Stewardship at UNC*. http://sils.unc.edu/sites/default/files/general/research/UNC_Research_Data_Stewardship_Report.pdf.

Ramírez, Marisa L. 2011. "Whose Role is it Anyway?: A Library Practitioner's Appraisal of the Digital Data Deluge." *Bulletin of the American Society for Information Science & Technology* 37 (5): 21-23. doi:10.1002/bult.2011.1720370508.

Westra, Brian, Marisa Ramirez, Susan Wells Parham, and Jeanine Marie Scaramozzino. 2010. "Selected Internet Resources on Digital Research Data Curation." *Issues in Science and Technology Librarianship* 63. http://www.istl.org/10-fall/internet2.html.

Witt, Michael. 2008. "Institutional Repositories and Research Data Curation in a Distributed Environment." *Library Trends* 57 (2): 191-201. doi:10.1353/lib.0.0029

Bibliography

Carnegie Foundation for the Advancement of Teaching. "Carnegie Classifications", accessed November 21, 2011, http://classifications.carnegiefoundation.org/descriptions/basic.php.

Cragin, M.,H., C. Palmer L., M. Kogan, J. Carlson R., and M. Witt. 2010. "Data Sharing, Small Science, and Institutional Repositories." *Philosophical Transactions of the Royal Society A* 368 (1926): 4023-4038. doi:10.1098/rsta.2010.0165.

Davis, Hilary M. and John N. Vickery. 2007. "Datasets, a Shift in the Currency of Scholarly Communication: Implications for Library Collections and Acquisitions." *Serials Review* 33 (1): 26-32. doi:10.1016/j.serrev.2006.11.004.

Gold, Anna. 2010. "Data Curation and Libraries: Short-Term Developments, Long-Term Prospects." *Office of the Dean (Library)*. http://digitalcommons.calpoly.edu/lib_dean/27.

Harvey, Ross. 2006. "Appraisal and Selection." In DCC Digital Curation Manual, edited by Seamus Ross and Michael Day: HATII, University of Glasgow; University of Edinburgh; UKOLN, University of Bath; Council for the Central Laboratory of the Research Councils. http://www.dcc.ac.uk/resource/curation-manual/chapters/appraisal-and-selection/.

Heidorn, P. B. 2011. "The Emerging Role of Libraries in Data Curation and E-Science." *Journal of Library Administration* 51 (7): 662-672. doi:10.1080/01930826.2011.601269.

James Madison University Office of Sponsored Programs. 2011. "Funding News & Notes." http://www.jmu.edu/sponsprog/FNN2011.pdf.

Markey, Karen, Beth St. Jean, Soo Young Rieh, Elizabeth Yakel, and Jihyun Kim. 2008. "Institutional Repositories: The Experience of Master's and Baccalaureate Institutions." Portal: Libraries and the Academy 8 (2): 157-173. http://muse.jhu.edu/journals/portal_libraries_and_the_academy/v008/8.2markey.pdf.

Nykanen, Melissa. 2011. "Institutional Repositories at Small Institutions in America: Some Current Trends." *Journal of Electronic Resources Librarianship* 23 (1): 1-19. doi:10.1080/1941126X.2011.551089.

Shreeves, Sarah L. and Melissa H. Cragin. 2008. "Introduction: Institutional Repositories: Current State and Future." *Library Trends* 57 (2): 89-97. doi:10.1353/lib.0.0037.

The National Science Foundation. 2010. *Proposal and Award Policies and Procedures Guide, Part II - Award and Administration Guide.* National Science Foundation. http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_index.jsp

Witt, Michael, Jacob Carlson, D. Scott Brandt, and Melissa Cragin. 2009. "Constructing Data Curation Profiles." *The International Journal of Digital Curation* 4 (3): 93-103. http://www.ijdc.net/index.php/ijdc/article/view/137.