Networked Instruction for Research Data Curation Education: The CRADLE Project

Helen Tibbo
School of Information & Library Science
University of North Carolina at Chapel Hill
201 Manning Hall, CB# 3360
Chapel Hill, NC 27599-3360
919-962-8063
tibbo@ils.unc.edu

Thu-Mai Christian
School of Information & Library Science
University of North Carolina at Chapel Hill
229C Davis Library, CB# 3355
Chapel Hill, NC 27599-3355
919-962-6293
thumai@email.unc.edu

ABSTRACT

In this paper, we describe a new initiative to develop a massive open online course (MOOC) for training library and information science students, library practitioners, and data producers in data curation. The Curating Research Assets and Data using Lifecycle Education (CRADLE) project exploits the affordances of MOOC technology to provide a networked learning environment that will encourage and foster the creation of research ecosystems in which CRADLE participants—library and information graduate students, library practitioners, and data producers—will have opportunities to collaborate with and learn from others engaged in data curation practice.

General Terms

communities, case studies and best practice, training and education

Keywords

data curation, data curation education, data management, massive open online course, MOOC

…data scientists [including] librarians [and] archivists… have the responsibility to design and implement education and outreach programs that make the benefits of data collections and digital information science available to the broadest possible range of researchers, educators, students, and the general public. – National Science Board, 2005

If data curation is viewed as a means to advance science … then libraries need to partner closely with investigators in the sciences and in other disciplines they serve. Because data vary so much by field, and by investigator, generic approaches to data collection are not feasible. – Christine Borgman, 2010

1. INTRODUCTION

While "standing on the shoulders of giants" and building on centuries of discoveries and painstaking research, much of 21st Century physical, medical, and social sciences are radically different from their predecessors that revolved around observation, experimentation, and more recently, small-scale computation. Today's "e-Science" (Hey & Hey, 2006) or "data-intensive science" (Gray & Szalay, 2007) or what Jim Gray of Microsoft Research termed in 2007 "fourth paradigm" science, (Bell, Hey, and Szalay, 2009; Gray, 2007; Hey, Tansley, and Tolle, 2009; Microsoft Research, 2006) is increasingly "carried out through distributed global collaborations enabled by the Internet" (UKNESC, 2012). This science features use and significantly, re-use, of very large data collections, very large scale computing resources, and high performance visualizations (Borgman, 2007; Borgman, 2012; Carlson & Anderson, 2007; SCARP Project, 2009). The stakes are high as e-Science promises discoveries and benefits not possible with more traditional methodologies. Social scientists are also facing the challenges of large-scale data. King observes that the "massive increases in the availability of informative social science data are making dramatic progress possible in analyzing, understanding, and addressing many major societal problems. Yet the same forces pose severe challenges to the scientific infrastructure supporting data sharing, data management, informatics, statistical methodology, and research ethics and policy, and these are collectively holding back progress" (King, 2011, p. 719). Humanists have also taken up the data-intensive approach, and the term "cyberscholarship" refers to scholarly research using high performance computing and digital libraries (American Council of Learned Societies, 2006; Arms, 2008).

Despite the apparent focus on technology, today's research environment is not just about high-capacity networks and large-scale digital data storage. It is not just about creating terabytes of new data or analyzing arrays of existing data in new ways. Effective and efficient data lifecycle management lies at the heart of today's research enterprise (DCC, "What"; Lord, Macdonald, Lyon, & Giaretta, 2004). For example, if data are not adequately or accurately described using metadata they will not be found in data stores, be interoperable, or understood for re-use. If sensitive data are not de-identified or kept securely, privacy and confidentiality will be breached. Data-intensive science presents a wide array of data management challenges for researchers, information and computer scientists, librarians, and data archivists as well as universities and public and private research laboratories that create and house data (ARL, 2007; Borgman, 2008; Choudhury, 2008; Garritano & Carlson, 2009; Gold, 2007a; Gold, 2007b; Hey & Hey, 2006; Jones, 2008). For truly productive science and scholarship that maximizes every research dollar and makes the investment in data creation re-usable, researchers must work in concert with data managers and digital curators (Abbot, 2008; DCC, 2010; Joint, 2007; Swan & Brown, 2008; National Academy of Sciences, 2009).

As e-Science takes root, producing unprecedented volumes of data in various and novel data formats, associated research data management challenges have also propagated. These challenges have invaded the purview of library and information science (LIS) professionals who are being called upon to tend to them. Many believe that data curation aligns with both the library mission to collect and provide access to scholarly materials and the librarian expertise that includes metadata, archival preservation, and bibliographic citation—all of which are applicable to data curation (Shaffer, 2013; Harris-Pierce & Liu, 2012; Latham & Poe, 2012). Others, however, argue that data curation necessarily restructures library practices because of the incongruences between the level and type of technical skill and professional judgment required for dealing with data and that required for other types of library materials (Gold, 2007a; Salo, 2010).

These incongruences, according to Gold (2007b), are resolved when libraries gain "fluency across library and scientific cultures" (Building capacity and understanding, para. 2). Consequently, LIS graduate schools have developed data curation education programs to teach such fluency. These programs not only teach data curation concepts such as digital preservation and metadata, but also they recognize that students benefit most from learning these concepts within the context of the research communities that produce data. The data curation specialization offered by the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign requires foundational courses based on the concept of *purposeful curation* that emphasizes the cultural context, unique characteristics, and frameworks of data production, management, and sharing, while also placing emphasis on practical field experiences (Palmer, Weber, Munoz, & Renear, 2013). Likewise, the University of North Carolina at Chapel Hill's School of Information and Library Science offers a post-master's certificate in Data Curation that requires students to complete independent study projects that give students practical experience in a work environment (University of North Carolina, 2014). Carlson et al. (2011) maintain the need for data information literacy (DIL) programs that give students the ability to interpret and analyze data beyond simply managing them, with course content grounded in the cultures and practices of disciplinary domains.

Our assertion is that data curation education programs need to go a step further to address e-Science trends that have obliged the scientific community to (re-)define cultures and practices around data production, management, and sharing (Gray, 2009). The abundance of data production, decentralization of data sources, and interdisciplinary collaboration have necessitated the development of new technological approaches to data management and dissemination that enhance knowledge sharing in the data-intensive research landscape (Bell, Hey, & Szalay, 2009; Gray, 2009; Hey & Trefethen, 2003). If data curation education programs are to remain responsive to the rapidly evolving needs of the scientific community, programs need to adopt parallel approaches for the training and mobilization of LIS professionals who will be expected to apprehend the context in which research data are produced, managed, and disseminated. Therefore, data curation education must not only teach students the requisite data curation concepts defined in established graduate curricula, but also they must situate students within the relevant contextual framework.

2. THE CRADLE PROJECT

The IMLS-funded *Curating Research Assets and Data using Lifecycle Education* (CRADLE) project is working to take this step by developing a massive open online course (MOOC) that will provide instruction on data curation principles while focusing squarely on learning through networks of data management education and practice. A noteworthy outcome of e-Science has been the creation of "research ecosystems" that exploit advances in Internet communications technology (Goodman & Wong, 2009). These research ecosystems have given the citizen scientist opportunities to make important contributions to the corpus of scientific discovery, offered flexibility that has enabled interdisciplinary collaborations for solving large-scale problems, and provided access to tools that make scientific data comprehensible to a broader audience of individuals with varying levels of expertise (Goodman & Wong, 2009).

Likewise, the MOOC platform will allow CRADLE participants to exploit the same technological affordances to promote and support learning in a networked environment. Learners will be given access to the necessary technology and tools to enable them to construct similar research ecosystems in which individuals will be able to engage with and learn from others involved in data curation practice and make contributions to greater discussions around data curation. CRADLE will not only teach librarians the skills required for preparing data for long-term preservation and use, but also foster knowledge ecosystems by:

- Assigning projects that require students to make contact with data producers and information professionals at their local universities, libraries, research centers, or data repositories;
- Hosting virtual summits for CRADLE graduates that provide ongoing opportunities to share data management experiences and continue engagement with data management issues;
- Sponsoring opportunities for CRADLE students and graduates to attend data management symposia that feature significant players across the data management landscape; and
- Establishing virtual sandboxes and other technology that enable students to collaborate on data management challenges, with each student assigned to different data management stakeholder roles.

While individual CRADLE learning modules on data curation topics will contain content aimed toward specific audiences—LIS students, library practitioners, and data producers—each type of individual will interact with one another to solve data management problems. These interactions will encourage and foster an environment in which they can seed networks, which will grow as students also engage with their local research communities to explore first-hand the challenges of data management.

Moreover, CRADLE will provide an environment that will aid in the alignment of efforts to promote standards of data curation practice and to shift the culture toward one that recognizes research data as valued assets essential to the sustainability of the research enterprise. Where LIS students, library practitioners, and data producers coalesce on solutions to data management problems, discoveries of commonalities in data culture and practices may inform the establishment of best practices and encourage their adoption. CRADLE will serve as the backdrop from which effective data management education and best practice will emerge. The dynamic and unpredictable nature of research in the *fourth paradigm* (Gray, 2009) requires a more profound engagement with the research community to allow data curation education to adapt accordingly. No longer can information professionals operate within the confines of deep-seated archival principles and practices; librarians and archivists must find station within research ecosystems populated by data management stakeholders.

3. CONCLUSION

As current programs and novel initiatives such as CRADLE continue to develop and evolve, further study will be necessary to determine their success in preparing the next generation of librarians and information professionals as well as researchers themselves in meeting data management requirements of funders, journal publishers and institutions. If and when these data curation programs are proven successful, “working with data will become a mature component of librarianship when it is accepted into regular library practices; when terms like ‘data reference’ become simply ‘reference’ and datasets are not given any more specific or specialized treatment than other library collections” (Witt, 2012, p. 186). For this to happen, librarians must become active participants in the research community, making meaningful connections to individuals confronting data challenges, and arriving at common solutions for overcoming those challenges.

4. ACKNOWLEDGMENTS

The CRADLE project is sponsored by the Institute of Museum and Library Services, under award #RE-06-13-0052-13

5. REFERENCES

[1] American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences. (2006). *Our cultural commonwealth* (No. The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences). New York, NY: American Council of Learned Societies. Retrieved from http://www.acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf

[2] Arms, W. Y. (2008). Cyberscholarship: High performance computing meets digital libraries. *The Journal of Electronic Publishing*, *11*(1). doi:10.3998/3336451.0011.103

[3] ARL Joint Task Force on Library Support for E-Science. (2007). Agenda for developing e-Science in research libraries (Final Report and Recommendations). Washington, D.C.: Association of Research Libraries Retrieved from http://old.arl.org/bm~doc/ARL_EScience_final.pdf

[4] Atkins, D. (2003). *Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure* (No. cise051203). Arlington, VA: National Science Foundation. Retrieved from http://www.nsf.gov/od/oci/reports/atkins.pdf

[5] Bell, G., Hey, T., & Szalay, A. (2009). Beyond the data deluge. *Science*, *323*(5919), 1297–1298. doi:10.1126/science.1170411

[6] Borgman, C. L. (2007). Scholarship in the digital age : information, infrastructure, and the Internet. Cambridge, Mass.: MIT Press.

[7] Borgman, C. L. (2008, June 27). *The role of librarians in e-science*. Conference Presentation presented at the European Conference of Medical and Health Libraries, Helsinki, Finland. Retrieved from http://blip.tv/eahil2008/the-role-of-libraries-in-e-science-christine-borgman-eahil2008-1045049

[8] Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078. doi:10.1002/asi.22634

[9] Carlson, S., & Anderson, B. (2007). What are data? The many kinds of data and their implications for data re-use. *Journal of Computer-Mediated Communication*, *12*(2). Retrieved from http://jcmc.indiana.edu/vol12/issue2/carlson.html

[10] Carlson, J., Fosmire, M., Miller, C. C., & Nelson, M. S. (2011). Determining data information literacy needs: A study of students and research faculty. *Portal: Libraries and the Academy*, *11*(2), 629–657.

[11] Choudhury, G. S. (2008). Case study in data curation at Johns Hopkins University. *Library Trends*, *57*(2), 211–220. doi:10.1353/lib.0.0028

[12] Digital Curation Centre (DCC). (n.d.). What is digital curation? Retrieved April 11, 2014, from http://www.dcc.ac.uk/digital-curation/what-digital-curation

[13] Digital Curation Centre (DCC). (2010). Resources for Digital Curators. *Introduction to Curation*. Retrieved from http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation

[14] Garritano, J., & Carlson, J. (2009). A subject librarian’s guide to collaborating on e-Science projects. *Issues in Science and Technology Librarianship*, *57*(Spring 2009). doi:10.5062/F42B8VZ3

[15] Gold, A. (2007a). Cyberinfrastructure, data, and libraries, part 1: A cyberinfrastructure primer for librarians. *D-Lib Magazine*, *13*(9/10). doi:10.1045/september20september-gold-pt1

[16] Gold, A. (2007b). Cyberinfrastructure, data, and libraries, part 2: Libraries and the data challenge: Roles and actions for libraries. *D-Lib Magazine*, *13*(9/10). doi:10.1045/september20september-gold-pt2

[17] Goodman, A., & Wong, C. (2009). Bringing the night sky closer: Discoveries in the data deluge. In T. Hey, S. Tansley, & K. Tolle (Eds.), *The fourth paradigm: Data-intensive scientific discovery* (pp. 39–44). Redmond, WA: Microsoft Research. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_part1_goodman_wong.pdf

[18] Gray, J. (2009). Jim Gray on e-Science. In A. J. G. Hey, S. Tansley, & K. M. Tolle (Eds.), *The fourth paradigm: Data-intensive scientific discovery*. Redmond, WA: Microsoft Research.

[19] Gray, J., & Szalay, A. (2007, January 11). *eScience--A transformed scientific method*. Presented at the Computer Science and Technology Board of the National Research Council, Mountain View, CA. Retrieved from http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt

[20] Harris-Pierce, R. L., & Liu, Y. Q. (2012). Is data curation education at library and information science schools in North America adequate? *New Library World*, *113*(11), 598–613. doi:10.1108/03074801211282957

[21] Hey, A. J. G., & Trefethen, A. (2003). The data deluge: An e-science perspective. In F. Berman, G. Fox, & A. J. G. Hey (Eds.), *Grid Computing: Making the Global Infrastructure a Reality* (pp. 809–824). New York: Wiley.

[22] Hey, T., & Hey, J. (2006). e-Science and its implications for the library community. *Library Hi Tech*, *24*(4), 515–528. doi:10.1108/07378830610715383

[23] Joint, N. (2007). Data preservation, the new science and the practitioner librarian. *Library Review*, *56*(6), 451–455. doi:10.1108/00242530710760337

[24] Jones, E. (2008). *e-Science talking points for ARL deals and directors*. Washington, D.C.: Association of Research Libraries. Retrieved from http://www.arl.org/storage/documents/publications/e-science-talking-points.pdf

[25] King, G. (2011). Ensuring the data-rich future of the social sciences. *Science*, *331*(6018), 719–721. doi:10.1126/science.1197872

[26] Latham, B., & Poe, J. W. (2012). The library as partner in university data curation: A case study in collaboration. *Journal of Web Librarianship*, *6*(4), 288–304. doi:10.1080/19322909.2012.729429

[27] Lord, P., Macdonald, A., Lyon, L., & Giaretta, D. (2004). From data deluge to data curation. In *Proceedings of the 3rd UK eScience All Hands Meeting* (pp. 371–375). Nottingham, UK: Citeseer. Retrieved from http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/150.pdf

[28] Microsoft Research. (2006). 2020 Science. Retrieved April 11, 2014, from http://research.microsoft.com/en-us/um/cambridge/projects/towards2020science/background_overview.htm

[29] National Academy of Sciences. (2009). Ensuring the integrity, accessibility, and stewardship of research data in the digital age. Washington, DC: The National Academies Press. Retrieved from http://www.nap.edu/catalog.php?record_id=12615#description

[30] Palmer, C. L., Weber, N. M., Munoz, T., & Renear, A. H. (2013). Foundations of data curation: The pedagogy and practice of “purposeful work” with research data. *Archive Journal*, (3). Retrieved from http://www.archivejournal.net/issue/3/archives-remixed/foundations-of-data-curation-the-pedagogy-and-practice-of-purposeful-work-with-research-data/

[31] Salo, D. (2010). Retooling libraries for the data challenge. Ariadne, 63. Retrieved from http://www.ariadne.ac.uk/issue64/salo

[32] SCARP Project. (2009). *Disciplinary approaches to sharing, curation, reuse and preservation* (Final Report). Bristol, UK: JISC. Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/scarp/SCARP-FinalReport-Final-SENT.pdf

[33] Shaffer, C. (2013). The role of the library in the research enterprise. *Journal of eScience Librarianship*, *2*(1), 8–15. doi:10.7191/jeslib.2013.1043

[34] Swan, A., & Brown, S. (2008). The skills, role and career structure of data scientists and curators: An assessment of current practice and future needs. UK: JISC. Retrieved from http://ie-repository.jisc.ac.uk/245/1/DataSkillsReport.doc

[35] United Kingdom National e-Science Centre (UKNESC). (2012). Defining e-Science. Retrieved April 11, 2014, from http://www.nesc.ac.uk/nesc/define.html

[36] University of North Carolina (2014). Post master’s certificate: Data Curation. *UNC School of Information and Library Science*. Retrieved from http://sils.unc.edu/programs/graduate/post-masters-certificates/data-curation

[37] Witt, M. (2012). Co-designing, co-developing, and co-implementing an institutional data repository service. *Journal of Library Administration*, *52*(2), 172–188. doi:10.1080/01930826.2012.655607