The International Journal of Digital Curation
Volume 8, Issue 1 | 2013

Model Development for Scientific Data Curation Education

Karon Kelly, Mary Marlino and Matthew S. Mayernik,
National Center for Atmospheric Research

Suzie Allard and Carol Tenopir,
University of Tennessee School of Information Sciences

Carole L. Palmer and Virgil E. Varvel Jr.,
Graduate School of Library and Information Science,
University of Illinois Urbana-Champaign

Abstract

The mounting and critical need for scientific data curation professionals was the impetus for the Data Curation Education in Research Centers (DCERC) program. DCERC is developing a sustainable and transferable model for educating Library and Information Science (LIS) students in data curation through field experiences in research and data centers. DCERC has established and implemented a graduate research and education program bringing students into the real world of scientific data curation, where they engage with current practices and challenges, and share their developing expertise and research. The DCERC partner institutions are developing and evaluating this model with the intention of scaling the program to a larger cadre of partners and participants. This paper reports on progress in the early phases of the model development.

International Journal of Digital Curation (2013), *8*(1), 255–264. http://dx.doi.org/10.2218/ijdc.v8i1.258

 doi:10.2218/ijdc.v8i1.258

Introduction

Data are now widely recognized as the currency of science, and "big data is a big deal," as exemplified by the U.S. Office of Science and Technology Policy announcement of the Big Data Research and Development Initiative (Kalil, 2012). Big data is the next frontier, expected to lead to significant economic growth and innovation (Manyika et al., 2011). As noted by Lucy Nowell from the U.S. Department of Energy at the 2010 Research Data Workforce Summit (Varvel et al., 2011), there are now forceful arguments for the need for curation services to scale up to meet the state of rapid growth in both data and computing (Shoshani & Roten, 2010). Science at large is in the very early stages of systematic approaches to data curation and management. While some research centers have developed systems and services to support these goals, most researchers themselves have little or no formal training in data management practices, and many express concern about their lack of preparedness in this area (Jahnke, Asher & Keralis, 2012). The conduct of data-intensive research in universities and research centers will increasingly rely on the skills and expertise of a relatively new and sophisticated professional cohort (Jahnke, Asher & Keralis, 2012). Broader collaboration among educators, researchers and practitioners is essential to fostering the emerging curation field (Botticelli et al., 2011).

Nowhere is this need more acute than in the Earth System Sciences. As a global community, we face rapidly escalating challenges arising from our large-scale impacts on the Earth system and our growing vulnerability to sudden and gradual systemic changes in the Earth system. Understanding the impacts of changing climate, the health effects of pollution, and the devastation wrought by extreme weather events are but a few examples of how data-driven science is critical to knowledge of our world. Appropriately curated data is central to informing science policy on national and international scales. Collaboration with exemplar data centers is likely to be the most direct path for effectively moving science toward "the best stewardship of data" through "disciplinary engagement with preservation institutions" (Lynch, 2008). There is a clear need for data curation practitioners to become more engaged with the formal education process – as instructors and mentors – to bring current issues and experiences to professional education (Varvel et al., 2011).

Responding to these challenges requires a capable, diverse scientific workforce to generate new knowledge. Equally important is a well-trained scientific data curation workforce that is capable of solving problems in an increasingly complex, data-driven world. Both in the United States and abroad, Library and Information Science schools are increasingly concerned with developing curricula to meet the growing need for well-trained professionals to provide services and tools for the curation, sharing and preservation of scientific research data (Interagency Working Group on Digital Data, 2009). Concurrent with rising awareness among information science programs about future workforce issues is the growing recognition within data centers of the need to improve their services for scientists and other data users, as well as a desire for better relationships with data management and curation educators, and domain-related data experts. Stakeholders who participated in the 2011 Research Data Workforce Summit recommended partnering education programs with data centers to facilitate internships and field experiences for students that include mentoring by data professionals, and

training programs that are integrative and general, yet allow for development of specialized expertise (Varvel et al., 2011).

The DCERC Model

It is clear that sustainable and transferable educational models for this new field of scientific data curation are past due. To address this need, we developed the Data Curation Education in Research Centers (DCERC) program, a collaboration of the Center for Informatics Research in Science and Scholarship (CIRSS) at the University of Illinois Urbana-Champaign Graduate School of Library and Information Science (Illinois GSLIS), the University of Tennessee School of Information Sciences (Tennessee SIS), and the National Center for Atmospheric Research (NCAR) to develop a model for educating Library and Information Science master's and doctoral students in data curation through core data curation coursework and field experiences in research and data centers. DCERC is funded through the Laura Bush 21st Century Librarians Program at the Institute of Museum and Library Services (IMLS), an independent US federal agency that funds research and education for library and museum professionals. Dr. Carole Palmer, professor and director of the CIRSS, serves as principal investigator of the DCERC program.

The first phase of DCERC has been built on the well-established strengths of these partners. The Illinois GSLIS and the Tennessee SIS contributed their existing base of education and research in data curation. NCAR provides internship/field experiences for library and information science students that are unprecedented in their design. The focus of the student internships is to provide students with strong mentoring and first hand experiences in the issues, methods and challenges of scientific data curation in scientific research contexts. The intent is to align these experiences as closely as possible with each student's disciplinary and research interests.

NCAR's contribution to the program is a continuation of its lengthy history of education, professional development, and mentoring and builds on successful internship programs with both undergraduate and graduate students. The significance of this mentoring activity is recognized through formal performance review and an institution-wide mentoring recognition award, peer reviewed in a highly competitive process. NCAR represents a state-of-the-art research center that has developed policies and practices that implement community recommendations to promote recognition of data workforce education and research within institutional reward structures (Varvel et al., 2011).

DCERC is being piloted by a cadre of six recruited students at Illinois GSLIS and Tennessee SIS. Master's students took a core course through the Illinois GSLIS online program, and will participate in two summer internships at NCAR, the first of which took place in 2012. PhD students will intern at NCAR during the 2013-2014 academic year.

An important and highly valued aspect of the program has been the development of DCERC student culture, establishing a strong cohort that provides students with a sense of identity, a community of learners, and the beginnings of a strong professional network. A cross-institutional forum facilitates the distribution of program information and resources on data curation for the students; helps in forming the

 doi:10.2218/ijdc.v8i1.258

social and scholarly community among students within and across institutions; and aids in the collection of evaluation information as part of the ongoing assessment of the program. The integration of activities and communication across institutions has been positively noted by students. In either emails or personal conversations, both masters and PhD students have mentioned advantages of interactions across the institutions, such as fluid mixing of students into research groups, access to diverse perspectives, and exposure to how another institution operates, which has been important context as students begin pursuing opportunities for professional positions.

Curriculum

The core course, Foundations in Data Curation, is a major component of the DCERC program. This semester-long graduate course provides an overview of a broad range of theoretical and practical problems in data curation and examines issues including appraisal and selection, long-lived data collections, research lifecycles, workflows, levels of representation, metadata, and legal and intellectual property issues. For the DCERC project, the course has been revised and updated in preparation for its delivery to students at both institutions. Thirty students from both universities participated in the course that engaged faculty and data professionals in its delivery. Particular attention was paid to updating readings and bringing both humanities and science data curation aspects to the course. The course was very highly rated in the overall formal evaluation. A significant measure of success was the fact that the average self-perceived data curation knowledge among the students increased by 60%. Additional activities, such as “brown bags”, were developed to expose students to knowledge and experiences that enhance the classroom experiences. Topics ranged from technical (e.g. research topics, grant writing) to domain specific (e.g. new data preservation strategies) to professional (e.g. how to write a resume, interviewing tips).

Data Curation Workshop

The summer internship at NCAR began with a three-day workshop on data curation. The workshop was the first visit to NCAR for most of these students and was designed to help them assimilate and relate what they had learned in their courses to the operations and goals of scientific research centers. The aim included introducing them to or strengthening their understanding of the role of data centers in the scientific research process and in data curation. The workshop was also designed to build community among students, scientists, and data mentors. Panels of data managers and scientists included discussions of issues in data use and reuse by science mentors, and of data practices and how they compare across sites, from the perspective of data managers from NCAR, the National Snow and Ice Data Center, a Long Term Ecological Research (LTER) network site and the NASA Distributed Active Archive Center at the Oak Ridge National Laboratory. Other participants discussed issues related to data citation, metrics of data use, metadata, instituting data services and criteria for their evaluation. Student participation was central to the workshop and included presentations on their specific research interests followed by team meetings with their mentors to discuss their internship project topic, scope, and anticipated outcomes as well as to identify the resources and process requirements of their projects, such as acquiring human subject research/institutional review board approval for their projects. Finally, students and faculty participated in a field trip to the NCAR

Research Aviation Facility for a tour of the research aircraft and instrumentation, and a scientist presentation on real-time data collection and analysis tools while one of the aircraft was in flight collecting data.

In workshop evaluations, the tools of data management and the coordination required between scientists and data managers, while not explicitly part of the agenda, were listed as important learning outcomes by the participants. Other areas of learning cited included databases and data collection, data repositories, qualitative aspects of data; the nature of scientific research and workforce needs (Varvel, 2012a).

Summer Research Experiences

A unique and important aspect of DCERC is its strong, formal mentoring structure. Strong mentoring helps students succeed academically and professionally (Kardash, 2000; Seymour et al., 2004) and the formal structure ensures equal access to mentoring. Good mentors model professional success, transfer necessary skills, provide relevant resources, advise about career options, introduce students professionally, assist in career placement, and provide inspiration and personal support (NAS, 1997). The DCERC model is patterned after the Significant Opportunities in Atmospheric Research (SOARS) program at NCAR, now operating in its 17th year. SOARS has been widely recognized through formal and informal assessments as a highly successful program. A key factor to its success has been the multiple-mentor model engaging research, writing, peer and community mentors in the students experience (Laursen et al., 2010). Funded by federal agencies with commitments to increasing a more diverse workforce, SOARS aims to broaden participation in the atmospheric and related sciences through a multi-year undergraduate-to-graduate bridge program that is equal parts learning community, mentoring program, and research internship. SOARS activities are designed to engage, encourage, nurture and prepare students as they pursue a career in atmospheric or related science.

Recognizing the value and success of the SOARS model, the DCERC internships utilize a similar mentorship model including scientist and data mentors, as well as a young investigator as peer mentor and an institutional community mentor. The science mentor guides student research practice; the data mentor guides student development of data curation skills and knowledge; and the peer mentor, a PhD information scientist, provides general support to students as their projects progress. Together the peer and community mentors guide students in their integration in the program and provide information on institutional policies, other workplace issues, and career opportunities at NCAR and other data centers and universities. Within this academic and mentoring structure, students receive intensive formal and practice-based instruction uniquely blended with real world experiences in a world-renowned data center.

Pairing of students and mentors was based on student's educational background, scientific interests, and level of experience in data curation and management. These were then matched with mentor disciplinary/domain expertise and project interests. Mentors participated in mentor training based on the SOARS program training. The interns worked on a variety of projects that ranged from investigating the social

 doi:10.2218/ijdc.v8i1.258

effects of climate change to studying data project workflows and determining ways to improve efficiency. These projects included:

- **Preparing data sets for ingest into a repository.** This project reviewed the data lifecycle from research to access, use, and reuse (Johns, 2012).
- **Evaluating climate model metadata.** This student explored ways data curators can collect metadata about context and use via two projects at NCAR (Thomer, 2012).
- **Auditing data management workflow.** This student reviewed varying data workflows of several groups within NCAR. The project investigated ways to increase efficiency and reduce redundancy (Eaker, 2012).
- **Assisting in data and metadata organization.** This project focused on helping both data managers and scientists to enhance communication, with the aim of creating a manual for best practice and a data management plan (Siddell, 2012).

The culminating activity of the summer internship for students, mentors and faculty was a poster session and keynote presentation by Dr. Christine Borgman, presidential chair and professor of Information Studies at the University of California, Los Angeles (UCLA), on issues and incentives for sharing research data. Students presented their posters reporting on their internships and discussed their projects with attendees. In addition, the DCERC student posters were accepted for formal presentation at the 8th International Digital Curation Conference.

Evaluation

In order to evaluate the effectiveness of the DCERC program, assess the impact of internships within a data curation curriculum, and maintain accountability for the funding, a multi-faceted evaluation and impact protocol was developed (Varvel, 2012b). An independent analysis of student exit interviews after the first summer experience indicates that students highly valued their mentors and the internship, felt that it positively affected their career plans, and that participation in this program had a major impact on their career decisions. Students enthusiastically agreed that the summer internship had substantially enhanced the impact of their overall education program by clarifying data curation in practical settings, demonstrating how the field is changing and helping them focus on what they wanted out of a career in data curation. For some, this meant considering a different career route given their increased awareness of jobs available in the field of data curation outside of academia. Other students stated more personal impacts, such as having a better understanding of both the challenges and opportunities presented by this emerging field. Because of the summer experience, one intern felt that it provided additional motivation to pursue jobs in data centers and develop contacts for future career options. At the opposite spectrum, one student cited that the experience reinforced the desire to pursue a career in a university library rather than a science data center. Perhaps most significant is that the majority of students felt more confident about their future in the field of data curation as result of participating in this program.

NCAR's experience with the program has been extremely positive from the perspective of both the mentors who worked with the interns and the senior leadership. It was anticipated that these internships would provide learning opportunities for mentors as well as students. Scientists and data mentors alike reported on the positive impact of their interactions with students as developing information professionals and new appreciation for the contributions they can make to issues of data management and curation. For example, one data mentor stated:

> "Explaining my work revealed various shortcomings and inadequacies in the processes involved, which will be improved and therefore make my work more efficient and productive."

Another noted:

> "I learned how the new generation of data curators think, design solutions, and how they perceive the expected audience of their work."

There was clear reciprocity in learning, with students providing direct support and best practices for curation, and NCAR science and data mentors providing their deep expertise and guidance on the real day-to-day challenges at a mature and sophisticated data operation (Palmer, 2012).

The outcomes and evaluation to date provide clear indicators on how to extend and enrich the DCERC program. We believe that we have developed a viable and sustainable model for consideration in other scientific disciplines, and one that provides a basis for abstracting best practices and contributing to a growing body of data curation education and practice research. Specifically, we anticipate that this model will provide a vehicle to:

- Increase student preparedness through hands-on experience and mentoring with seasoned data professionals;
- Promote greater understanding of the importance of digital curation within scientific disciplines;
- Build bridges between information schools and constituencies who are in need of data curation services.

Conclusion

The goal of DCERC is to prepare the next generation of leaders in scientific data curation. By offering strategic internship opportunities and mentoring for students, along with academic preparation, we have developed a unique program that is equal parts research internship, cohort development and mentoring program, and offers comprehensive multi-year support as students make the transition into a productive professional life. We will continue to refine and test this model to meet the needs of our future Earth System Science workforce, and have plans at NCAR to institutionalize the model by extending our existing graduate and post graduate fellowships to include data curation specialists. We are especially gratified that, in our case, the model proved to be a "win-win" for both the students and the science and

data mentors. We believe that this model may be transferable to many other disciplines, in both the sciences and the humanities, and we encourage others to build upon and refine our initial efforts.

Acknowledgements

We wish to express our sincere appreciation and acknowledge the contributions of the data and science mentors at NCAR:

Data Mentors:

- Bob Dattore, Computational and Information Systems Laboratory
- Scot Loehrer, Earth Observing Laboratory
- Gary Strand, NCAR Earth System Laboratory
- Steven Worley, Computational and Information Systems Laboratory

Science Mentors:

- Paty Romero Lankao, Research Applications Laboratory
- Samuel Levis, NCAR Earth System Laboratory
- David Schneider, NCAR Earth System Laboratory
- Dennis Shea, NCAR Earth System Laboratory
- Katy Young, Earth Observing Laboratory

References

Botticelli, P., Fulton, B., Pearce-Moses, R., Szuter, C. & Watters, P. (2011). Educating digital curators: Challenges and opportunities. *International Journal of Digital Curation 6*(2), 146-164. doi:10.2218/ijdc.v6i2.193

Eaker, C. (2012). *Data audit and analysis: Mapping the data workflow from ingest to archive.* DCERC Summer Internship. Boulder, CO: University Corporation for Atmospheric Research. Retrieved from http://opensky.library.ucar.edu/collections/OSGC-000-000-010-444

Interagency Working Group on Digital Data (IWGDD). (2009). *Harnessing the power of digital data for science and society.* Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council. Washington, DC: Office of Science and Technology Policy. Retrieved from http://www.nitrd.gov/About/Harnessing_Power_Web.pdf

Jahnke, L., Asher, A. & Keralis, S.D. (2012). *The problem of data.* Council on Library and Information resources (CLIR) Publication 154. Retrieved from http://www.clir.org/pubs/reports/pub154

Johns, E. (2012). *The data lifecycle flow: For me, this time.* Poster from the DCERC Summer Internship. Boulder, CO: University Corporation for Atmospheric Research. Retrieved from http://opensky.library.ucar.edu/collections/OSGC-000-000-010-445

Kalil, T. (2012). *Big data is a big deal* [blog post]. U.S. Office of Science and Technology Policy (OSTP), Executive Office of the President. Retrieved from http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal

Kardash, C.M. (2000). Evaluation of an undergraduate research experience: Perceptions of undergraduate interns and their faculty mentors. *Journal of Educational Psychology, 92*, 191-201. doi:10.1037/0022-0663.92.1.191

Laursen, S.L., Hunter, A., Seymour, E., Thiry, H. & Melton, G. (2010). *Undergraduate research in the sciences: Engaging students in real science.* San Francisco, CA: Jossey-Bass.

Lynch, C. (2008). Big data: How do your data grow? *Nature 455*, 28-29. doi:10.1038/455028a

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. & Byers, A.H. (2011). *Big data: The next frontier for innovation, competition, and productivity.* McKinsey Global Institute. Retrieved from http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation

National Academy of Sciences. (1997). *Adviser, teacher, role model, friend: On being a mentor to students in science and engineering.* Washington D.C.: National Academies Press. Retrieved from http://www.nap.edu/openbook.php?record_id=5789

Palmer, C.L. (2012). *DCERC proves to be a successful model for data curation education through field experience, curriculum.* Urbana-Champaign, IL: University of Illinois. Retrieved from http://cirssweb.lis.illinois.edu/DCERC/DCERCVideo2.html

Seymour, E., Hunter, A.B., Laursen, S.L. & DeAntoni, T. (2004). Establishing the benefits of research experiences for undergraduates in the sciences: First findings from a three-year study. *Science Education*, *88*, 493-534. doi:10.1002/sce.10131

Shoshani, S. & Roten, D. (2010). *Scientific data management: Challenges, existing technology, and deployment.* Boca Raton, FL: CRC Press

Siddell, K.D. (2012). *Now you are speaking my language: Translating and facilitating between researchers and data managers.* Poster from the DCERC Summer Internship. Boulder, CO: University Corporation for Atmospheric Research. Retrieved from http://nldr.library.ucar.edu/repository/collections/OSGC-000-000-011-152

Thomer, A. (2012). *Curating context and use: Pulling scientific workflows into the repository.* Poster from the DCERC Summer Internship. Boulder, CO: University Corporation for Atmospheric Research. Retrieved from http://nldr.library.ucar.edu/repository/collections/OSGC-000-000-011-528

Varvel, V.E. Jr. (2012a). *DCERC data curation workshop evaluation.* Champaign, IL: Center for Informatics Research in Science and Scholarship. Retrieved from http://hdl.handle.net/2142/35296

Varvel, V.E. Jr. (2012b). *DCERC data curation internship evaluation.* Champaign, IL: Center for Informatics Research in Science and Scholarship. Retrieved from http://hdl.handle.net/2142/35297

Varvel, V.E. Jr., Palmer, C.L., Chao, T. & Sacchi, S. (2011). *Report from the Research Data Workforce Summit.* Champaign, IL: Center for Informatics Research in Science and Scholarship. Retrieved from https://www.ideals.illinois.edu/handle/2142/25830