

**Data Curation Education in Research Centers**

Carole L. Palmer, Ph.D.
Center for Informatics Research in Science & Scholarship
Graduate School of Library and Information Science
University of Illinois at U-C
501 E. Daniel Street
Champaign, IL 61820-6212
1 (217) 244-3302
cpalmer@illinois.edu

Suzie Allard, Ph.D.
School of Information Sciences
University of Tennessee
453 Communication Building
Knoxville, TN 37996
1 (865) 974-1369
sallard@utk.edu

Mary Marlino, Ed.D.
National Center for Atmospheric Research Library
1850 Table Mesa Drive
Boulder, CO 80303-3000
1 (303) 497-8350
marlino@ucar.edu

**ABSTRACT**

New data skills are critical to the progress of 21st century science to ensure that data are properly selected and stored and can be readily discovered, accessed, and used over time. The Data Curation Education in Research Centers (DCERC) program will establish a model for data curation graduate education that enriches students learning and expertise as they conduct research and develop as professionals through field experience in a data intensive scientific environment. DCERC is a joint effort between two LIS schools and a prominent national scientific research center.

2. **THE SCIENTIFIC CONTEXT**

In the digital age, data-intensive science, which applies data-driven approaches to generate new information and insights [3], has become commonplace. Sophisticated observational instruments and computational modeling techniques generate massive amounts of data at varying scales (or levels) of analysis, requiring petabyte-scale computing. The resulting “digital deluge” [4], often associated with “big science,” such as particle physics (e.g., Large Hadron Collider), astronomy (e.g., the Sloan Sky Survey), and atmospheric science (e.g., coupled climate models), is impacting a range of scientific disciplines that increasingly rely on computational uses of data, including bioinformatics, crystallography, and engineering design [5].

The grand scientific challenges of our time also require collaborative and interdisciplinary scientific approaches, adding additional layers of complexity to the data curation requirements. The issues are not simply technical but also socio-cultural, economic, ethical, and political. The recent case of security leaks and theft of climate change data in the UK exemplifies how these non-technical dimensions play into the scientific process in general, but also the particular centrality of data in the functioning of the scientific enterprise [6].

3. **THE DATA WORKFORCE PROBLEM**

New data skills are critical to the progress of 21st century science to ensure that data are properly selected and stored, can be readily discovered, accessed, and used, and have their logical and physical integrity maintained over time [7]. The importance of the emerging profession of data curation is now widely recognized by universities, scientific organizations, and funding agencies [8,9]. As a newly established area of professional expertise, the workforce shortage is coupled with a lack of established best practices to provide a foundation for training in data management, archiving, and preservation—core elements of data curation. A solid base of information science research focused on data curation is essential for developing best practices and curriculum for education programs [10, 11]. The DCERC program is designed to build capacity in both the professional and research workforce, with a strong focus on research and development of best practices in the emerging field.
4. RESPONSE IN LIS EDUCATION
In the United States, research librarianship has begun to embrace its professional role in the distributed network of data collections and services [12, 13, 14]. The Institute of Museum and Library Services’ (IMLS) Laura Bush 21st Century Librarian Program, is supporting a significant range of efforts that have positioned Library and Information Science (LIS) schools as leaders in data curation education, including projects at Syracuse University (2009), University of Arizona (2006), University of Michigan (2008), and University of North Carolina (2006, 2008). DCERC is built on an existing foundation of IMLS projects at two of the collaborating institutions: an established and growing masters specialization, which began in 2006 for the sciences, extended to the humanities in 2008; and a program begun in 2004 establishing a masters science librarian specialization, and a doctoral program training LIS educators in science data and information.

5. KEY ELEMENTS OF DCERC
5.1 Training of researchers and professionals
The goal of DCERC is to develop a sustainable and transferable model for educating LIS masters and doctoral students in data curation through field experiences in research and data centers. DCERC brings students into the real world of scientific data curation, where they will engage with current practices and challenges in data collection in the field, as well as delivery, management, and preservation of a variety of observation, model, and simulation data at a research center, where students will have both domain science and data science mentors.

Target areas for doctoral student dissertation research include:
- cross-disciplinary data sharing and reuse potentials
- ontology of datasets, formats, provenance, identity conditions
- metadata for description, discovery, interpretation, integration
- interoperability, provenance, preservation, and reuse
- research data in the scholarly communication continuum
- trust, security, confidentiality, ownership, quality, attribution

Masters students are expected to gain a range of skills including:
- understanding of clients' information needs and content
- ability to critically evaluate, select, and filter data resources
- ability to find, evaluate, and synthesize relevant data sources
- ability to manage many aspects of the data lifecycle
- understanding of how to manage the diversity, size, and complexity of current and future data sets.

5.2 Breadth of impact
The program will have broad, demonstrable impact for education in the field by providing:
- a model for LIS schools and science organization partnerships
- a core of experienced graduates to assume leadership positions
- enhanced curation curriculum to be shared with other schools
- diverse, top-level students for the profession
- future LIS faculty with deep data curation expertise
- a demonstration of key roles of LIS in e-science
- collaborations focused on scientific problems
- explication of curation best practices

5.3 Alignment with NSF DataNet Projects
DCERC will also leverage other exemplary efforts where LIS professionals and researchers are making substantial contributions to data curation, particularly two current DataNet projects funded by the National Science Foundation. DataNet integrates “library and archival sciences, cyberinfrastructure, computer and information sciences, and domain science expertise to provide long-term preservation, access, integration, and analysis capabilities for digital data” [15]. Engagement with the DataNet projects will allow students to interact with active R & D groups and exposure them to early advances and results.

6. CONCLUSION
By increasing accessibility and alleviating some of the very real challenges inherent in the archiving, use, and reuse of large, complex data sets, scientists and researchers will be better able to conduct the vital research necessary for solving grand challenges in science. This progress, and future efforts like DataNet, however, will not be possible without an expert workforce in both data curation research and practice. DCERC is an important contribution in LIS education toward this goal.

6. ACKNOWLEDGMENTS
DCERC is supported by IMLS award # RE-02-10-0004-10.

6. REFERENCES


