

Centuries of Knowledge  
Data Curation Education Program

IMLS RE-05- 05-0036

University of Illinois at Urbana-Champaign

Annual Report #1  
October 1, 2006 – October 31, 2007

P. Bryan Heidorn, Ph.D., PI  
Graduate School of Library and Information Science  
(217) 244-7792  
[pheidorn@uiuc.edu](mailto:pheidorn@uiuc.edu)

Melissa Cragin  
Project Coordinator  
(217) 244-5574  
[cragin@uiuc.edu](mailto:cragin@uiuc.edu)

Ellen Rubenstein  
Research Assistant  
(217) 244-5574  
[erubens3@uiuc.edu](mailto:erubens3@uiuc.edu)

## **Performance Description**

Please address the following questions and requests for information related to the progress achieved on the project during the reporting period.

### **a. What is the purpose of the project?**

The primary goal of the Data Curation Education Program (DCEP) is to design a program of graduate study that can serve as a model for training data curators (DCs) within the context of a larger LIS education. Secondly, we intend to integrate this graduate training with ongoing research and practice to produce specialists that understand the research culture and can make substantive contributions to the mission of scientific, humanities, social science, and cultural heritage institutions and libraries.

We are developing a program to train a new generation of LIS professionals qualified as data curators and provide continuing education opportunities for practitioners already in the field. Through workshops, conference presentations, publications, and the influence of our graduates, it will also raise the visibility of the importance of information specialists in managing our knowledge resources for many decades, possibly centuries, to come. These students will become the leaders who build and maintain data systems to work in concert with the many digital libraries, archives, and repositories, as well as the indexing systems, metadata standards, ontologies, taxonomies, and vocabularies associated with digital data and products.

During the first year of the project, we have focused on work in several areas:

1. Organizing the project and coordination of activities
2. Establishing and working with our Advisory Committee
3. Curriculum development (including outreach for internship sites)
4. Recruitment of students
5. Needs Assessment
6. Planning for 2008 Summer Institute

Specific activities are described in the following sections, and organized by the objectives outlined in the grant proposal.

**b. What activities or services have been carried out with project funds to support the purpose of the project? If the project schedule has not been met, explain why and describe the steps being taken to return the project to its proposed schedule of completion.**

#### 1. Organizing the project and coordination of activities

Melissa Cragin was hired to be the half-time Project Coordinator, and started on October 1, 2006. Melissa is a doctoral candidate at the Graduate School of Library and Information Science at the UIUC.

- We have set up a project wiki, a tool we use for collaborative development of project plans and materials. The Data Curation Education Program wiki is not publicly accessible, but we would be happy to provide a password for IMLS access.
- A new website was set up at <http://www.uiuc.edu/goto/dcep> providing comprehensive coverage of the activities and products of the project, with links to the materials we have produced (such as papers and syllabi).
- The DCEP has been well publicized by our Publications & Communications office at GSLIS. In addition to giving papers at several conferences, we have made announcements at relevant conferences and meetings such as the Annual Meeting of the American Society for Information Science and Technology, the Digital Library Federation Forum, and the American Library Association.
- We are working now to develop a list of academic programs for direct contact and recruiting.
- Ellen Rubenstein was hired to be half-time Research Assistant, and started on August 16, 2007. Ellen is a doctoral student at the Graduate School of Library and Information Science at the UIUC. She has been working on developing lists of research and data centers to contact about setting up practicum and internship opportunities, as well as lists of relevant academic programs for student recruitment. Ellen will also help to plan and coordinate the Summer Institute for Data Curation, which is scheduled for June, 2008.

## 2. Advisory Committee

We had the first annual meeting of the DCEP Advisory Committee on February 16, 2007. The overarching goal of the advisory committee meetings is to learn what we need to teach LIS students to become professional data curators and to develop case studies and a set of best practices for teaching data curation expertise. The initial group was selected for coverage of a broad range of biological science; however, over the course of the DCEP the panel will be expanded to represent data curation issues from different disciplinary domains.

The current Advisory Committee includes:

- ◇ Thomas Garnett, Associate Director for Digital Library and Information Systems, Smithsonian Institution
- ◇ Gen. William D. Goran, US Army ERDC-CERL, Champaign, IL
- ◇ Katherine McNeill-Harman, Data Services and Economics Librarian, Massachusetts Institute of Technology
- ◇ Joanna McCaffrey, Collections Database Architect, The Field Museum
- ◇ Maryann Martone, Ph.D., Co-Director for the National Center for Microscopy and Imaging Research (NCMIR) University of California, San Diego

- ◇ Chuck Miller, Vice President, Information Technology and Chief Information Officer, Missouri Botanical Garden, St. Louis, MO
- ◇ Chris Rewerts, Ph.D., US Army ERDC-CERL Champaign, IL
- ◇ Indra Neil Sarkar, Ph.D., Informatics Manager, Marine Biological Laboratory, Woods Hole, MA
- ◇ Chris Freeland, Application Development Manager and Project Manager, Missouri Botanical Garden
- ◇ Martin Kalfatovic, Head of the New Media Office and Preservation Services Department, Smithsonian Institution Libraries, Washington, D.C.

Those attending the initial meeting included four scientists (research scientists and professors), two database managers / developers, two librarians, and the head of a library / data center. Development of a set of “core skills” will be informed by the committee’s experience and domain expertise, and their characterization of current data curation problems. The objectives set for the first meeting were to:

- describe a base of knowledge necessary to secure data-related jobs in research centers;
- develop a list of skills needed to carry out data management, curation, and archiving tasks for the next 3-5 years;
- identify requirements for internships at various sites.

During the meeting, the group detailed a range of skills needed that were particular to professional knowledge, domain knowledge, and personal competencies. Professional skills identified included knowledge and handling of various file formats, basic data file care, metadata, and use of applications across platforms. Domain-based knowledge included understanding research practices and scientific workflows, and the range of research problems in the domain and the relationship of techniques, instruments and data types used.

The advisory group also identified a number of personal skills and values necessary for success in data curation positions. These included maintenance of current awareness; analytical and problem solving skills; flexibility; ability to communicate with a variety of people; and a willingness to advocate for researchers’ participation. Finally, with regard to practicum opportunities and internships, the advisory group was eager to host students and expected that these placements could be mutually beneficial.

### **Ongoing Advisory Committee Work**

We had a second DCEP Advisory Committee meeting on October 1, 2007. We discussed the progress and curriculum of our Foundations of Data Curation class as well as the Digital Preservation class to be offered in spring, 2008. We outlined our plans for our Data Curation Summer Institute, where practicing academic librarians will be able increase their knowledge about data curation issues (see details below).

We further addressed internships and practica, noting that we have four sites ready for our students, including the Smithsonian, National Library of Medicine, Purdue University Library and Johns Hopkins Library. We discussed our recruitment efforts, as

well as our needs assessment. These activities are described in more detail below. We also developed several case study scenarios, and discussed how these could be most useful to students in the program.

Later this fall we will invite several individuals to join the Advisory Committee, in order to expand our local knowledge base, as well as to facilitate recruiting and internship opportunities. In March of 2008, we will hold our next (the third) meeting at the Missouri Botanical Garden in St. Louis, MO.

### 3. Curriculum Development

We have developed two new courses that are required in the Data Curation concentration: **Foundations in Data Curation**, and **Digital Preservation**. Students who participate in a practicum field placement will also be required to enroll in a Field Experience seminar. We have developed a list of recommended electives from which students will take 2-4; the list includes: Digital Libraries: Research and Practice; Museum Informatics; Electronic Publishing; Metadata in Theory and Practice; Ontology Development (Ontologies in the Natural Sciences OR Ontologies in Humanities); Information Transfer and Collaboration in Science; Design of Digitally Mediated Information Services; Interfaces to Information Systems; Information Modeling; Biodiversity and Ecoinformatics; Information Systems Analysis and Management.

Required courses for the DCEP concentration are offered through our online education option, LEEP, so that students without on-campus access to the University of Illinois at Urbana-Champaign will be able to complete the data curation concentration. In addition, the data curation courses will be available to students who are “undeclared” or outside of the DC program. Course materials will be available through the DCEP web site, as well as the University of Illinois IDEALS institutional repository.

### 4. Student Recruitment

We have compiled a recruitment mailing list consisting of:

- a. All undergraduate programs in biology, geology, geography, environmental sciences, and anthropology in Illinois
- b. McNair Scholars Program sites in selected colleges and universities throughout the United States
- c. Colleges and universities with undergraduate library and information science programs

In late fall we will be sending a letter, poster and brochures with information about the DCEP to these schools to inform them of our new program. We have also enlisted the aid of the Advisory Committee in this effort.

### 5. Needs Assessment

The first year Needs Assessment will be a survey of the approximately 400 researchers who make up the Faculty of the Environment at the University of Illinois. We are conducting this year's investigation collaboratively with the Environmental Council (EC) at the University of Illinois at Urbana-Champaign. Our two organizations are working together to understand how current data sets or collections are being used across UIUC, and the nature and extent of associated data management practices and problems. This survey is the first stage of a three-year assessment concerning data curation expertise, in which we will survey research departments, labs, and informatics initiatives and data centers across the country.

This campus-based survey will be followed by interviews with respondents who volunteer to provide more in-depth information on the value and use of environmental data collections, as well as how new data curation professionals can contribute to research operations and the management of valued data for long-term use. Survey and interview data will be used by the EC to develop data storage and support services. The faculty and staff on the DCEP project will use the data to inform curriculum planning and course development, as well as add to our knowledge about data curation.

At this time, we have completed the survey instrument pre-test, and the Pilot Study will be distributed this week (10/29). While we had originally planned to run the Pilot Study in the summer, the processes of developing the survey itself and conducting the pre-test took longer than we had anticipated. We expect to launch the survey to the full group by mid-December, 2007.

#### 6. 2008 Data Curation Summer Institute

During these last few months, practicing librarians have made requests to DCEP faculty and staff for training or classes in data curation. To address this need, we have begun planning for a 4-day Summer Institute on data curation, which will be held June 2-5 on the UIUC campus. This institute will be offered to 20-24 academic and research librarians who are involved in active work with research data in their university library. We will feature guest speakers with expertise in various aspects of data curation, including metadata and standards, preservation, intellectual property, appraisal and selection, and planning.

#### **c. What are the outputs of the project activities or services to support the purpose of the project? Explain what documentation is used to report the outputs.**

Materials developed *about* the DCEP and dissemination activities are listed on the project web site; the materials are also deposited into the UIUC institutional repository, the Illinois Digital Environment for Access to Learning and Scholarship (IDEALS). In addition, some materials will also be provided through the program or proceedings of conferences and workshops that we participate in. For example, we presented a paper in April, 2007, at the DigCCurr 2007 Symposium at the University of North Carolina, Chapel Hill. The program for the symposium is available on the DigCCurr website, with links to our paper:

[http://www.ils.unc.edu/digccurr2007/papers/heidornEtal\\_paper\\_8-2.pdf](http://www.ils.unc.edu/digccurr2007/papers/heidornEtal_paper_8-2.pdf)

**d. What are the outcomes of the project activities or services to support the purpose of the project? Explain what documentation is used to report the outcomes.**

We have presented at several conferences and workshops over the last several months, and have several upcoming conferences as well. Some of these were directly related to the DCEP, and others, while more peripheral, provided a suitable audience for introducing data curation and / or developing new collaborators or field experience opportunities.

Papers (Peer Reviewed)

Palmer, C.L., Heidorn, P.B., Wright, D. & Cragin, M.H. (in review). Graduate Curriculum for Biological Information Specialists: A Key to Integration of Scale in Biology. *International Journal of Digital Curation*.

Palmer, C.L., Cragin, M.H., Heidorn, P.B., Smith, L.C. (in review). Data Curation for the Long Tail of Science: The Case of Environmental Sciences. Submitted to the Third International Digital Curation Conference, Washington, DC, December 11-13, 2007.

Heidorn, P.B., Palmer, C.L., Cragin, M.H., & Smith, L.C. (2007). Data Curation Education and Biological Information Specialists. *DigCCurr2007: An international symposium on Digital Curation*, Chapel Hill, NC, April 18-20, 2007.

Cragin, M.H. (2006). The Roles of Shared Data Collections in Neuroscience. *Proceedings of the 2nd International Digital Curation Conference, "Digital Data Curation in Practice."* Glasgow, Scotland, November 21-22, 2006.

Heidorn, P.B., Palmer, C.L., Wright, D., & Cragin, M.H. (2006). Graduate Curriculum for Biological Information Specialists: A key to integration of Scale in Biology. *Proceedings of the 2nd International Digital Curation Conference, "Digital Data Curation in Practice."* Glasgow, Scotland, November 21-22, 2006.

Panels and Talks

Heidorn, P.B. Biological information management from molecules to ecosystems. Panel on "Biological Information Specialist Training," at 2nd International BioCuration Meeting, 2007 San Jose, California, October 25-28, 2007.

Heidorn, P.B. & Tibbo, H. (2007). Identifying Best Practices and Skills for Workforce Development in Data Curation. Panel at ASIS&T 2007 Annual Meeting, *Joining Research and Practice: Social Computing and Information Science*, Milwaukee, Wisconsin, October 19-24, 2007.

Cragin, Melissa H.; D'Avolio, Leonard; MacMullen, W. John; Smith, Catherine Arnott. (2007). The Effects of Context on Data Quality in Biomedical Data Reuse. Panel at the 2007 Annual Meeting of the American Society for Information Science & Technology (ASIS&T), Milwaukee, Wisconsin, October 19-24, 2007.

Cragin, M.H., MacMullen, W.J., Wallis, J., & Zimmerman, A. (2006). Scholarly communication: Scientists' views of a changing landscape. Panel titled, "Managing Scientific Data for Long-term Access and Use," Proceedings of the 69<sup>th</sup> Annual Meeting of the American Society for Information Science and Technology, ASIS&T '06, Austin, November 3-8, 2006.

### Posters

Heidorn, P. Bryan, Palmer, Carole L., Wright, Dan & Cragin, Melissa. Biological Information Specialist's Training. Poster for the 2nd International BioCuration Meeting, San Jose, California, October 25-28, 2007.

Heidorn, P. Bryan, Palmer, Carole L., Wright, Dan & Cragin, Melissa. Information Specialist's Training in Biology. Poster at Plant Science and Botany 2007 Conference, July 7-11, 2007, Chicago.

Cragin, M.H., Heidorn, P.B., Palmer, C.L., & Smith, L.C. (2007). An Educational Program on Data Curation. Poster Session at the Science and Technology Section program: Issues and Trends in Digital Repositories of Non-textual Information: Support for Research and Teaching. ALA Annual Conference, Washington, D.C., June 21-27, 2007.

### **e. Report other results of the project activities.**

P. Bryan Heidorn, Ph.D. (PI) co-chaired the LandInformatics Symposium and Workshops, September 7, 8, 28, 29, 2007.

P. Bryan Heidorn, Ph.D. hosted Arthur Chapman webcast: Principles and Methods of Biodiversity Data Quality, October 3, 2007.

Melissa Cragin (Project Coordinator) and Ellen Rubenstein (Research Assistant) participated in the "Shaping Outcomes" training from August 6–September 17, 2007.

Melissa Cragin is on the Program Committee for the Third International Digital Curation Conference, Washington, D.C., December 11-13, 2007.

Melissa is also co-editing an issue of *Library Trends* (2008) with Sarah Shreeves, the Coordinator for IDEALS, the institutional repository (IR) at the University of Illinois at Urbana-Champaign. This issue of *Library Trends* is focused on current research and practice for IRs. Melissa will specifically edit a section on IRs and models for managing research data.

Carole L. Palmer, Ph.D., Associate Professor at the Graduate School of Library and Information Science, UIUC, is attending two workshops:

- ◇ Workshop on Scientific and Scholarly Workflow Cyberinfrastructure, Baltimore, Maryland, Oct 3-5, sponsored by NSF and the Mellon Foundation.
- ◇ Workshop on Sharing and Curating Research Data, Washington, D.C. December 14, sponsored by the Mellon Foundation and JISC

In a collaboration with the Purdue University Libraries, we developed an IMLS National Leadership grant proposal titled, “Investigating Data Curation Profiles across Research Domains.” This application was successful, and we will soon begin a study across both the Purdue and UIUC campuses, with particular focus on disciplinary differences in research workflows and data practices, and their implications for data curation services. A part of this study will be to investigate ways librarians can interact with and support scientists in making their research output available.