

Centuries of Knowledge
Data Curation Education Program

IMLS RE-05- 05-0036

University of Illinois at Urbana-Champaign

Interim performance Report #2
November 1, 2007 – April 30, 2008

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

Melissa Cragin
Project Coordinator
(217) 244-5574
cragin@uiuc.edu

Ellen Rubenstein
Research Assistant
(217) 244-5574
erubens3@uiuc.edu

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 this last period of the project, we focused on work in several areas:

1. Curriculum development (including development and implementation of new required courses)
2. Recruitment of students
3. Development of internship opportunities
4. Needs Assessment
5. Working with our Advisory Committee
6. Planning for and developing the 2008 Summer Institute on Data Curation
7. Coordination and dissemination activities

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

Activities and services carried out with project funds during this reporting period:

1. Curriculum Development

We initially identified two new courses to develop and serve as the core requirements for the Data Curation concentration: **Foundations of Data Curation**, and **Digital Preservation**. The Foundations of Data Curation course was offered for the first time in the fall, 2007 semester, and was co-taught by Melissa Cragin and W. John MacMullen, a new Assistant Professor here at GSLIS. Enrollment for the course included 14 master's students, 1 CAS student, 3 Community Credit students, and 3 auditing. One of the Community Credit students has since applied to GSLIS MS program with the intention of completing the DC concentration. John MacMullen will teach the Foundations course again in the fall semester, 2008.

Digital Preservation was offered this semester (spring, 2008), and taught by Jerome McDonough, Assistant Professor at GSLIS. Enrollment for this class was 22, composed of 18 master's students, 2 CAS students, and 2 Community Credit students.

During this reporting period, we refined the requirements for the concentration, adding a third required course (to begin fall, 2008): **Systems Analysis and Management**. We have also reduced the list of recommended electives (from which students must select 2-4); the list includes: Foundations of

Information Processing in LIS; Biodiversity and Ecoinformatics; Digital Libraries: Research and Practice; Document Modeling; Information Modeling; Metadata in Theory and Practice; Ontology Development; and Representing and Organizing Information Resources.

Course materials, such as the syllabi for the Foundations of Data Curation course, will be made available through the DCEP web site, as well as the University of Illinois IDEALS institutional repository.

2. Student Recruitment

We developed a new flyer and poster for the program, both of which were professionally produced and designed for both recruiting and public relations purposes. These new materials are attractive, and feature colorful panels, photographs and text. The flyer provides an overview of the DC concentration, and includes required coursework, scheduling options, application procedures and financial aid information. The poster highlights the program in more general ways, by inviting readers to think about what will be required for digital information to remain accessible in the future.

We 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 mid-winter, we sent a letter, poster and flyers with information about the DCEP to 91 departments and programs to inform them of our new DC concentration. The McNair Scholars program was included on the recommendation of William Welburn, Associate Dean in the Graduate College at UIUC, as the McNair program supports African-American students, a traditionally underserved population from which we would like to draw GSLIS students. We have also enlisted the aid of the Advisory Committee in the general recruiting effort.

This past year we had 5 students enrolled in the program; we are expecting approximately 10 new students to enroll as of fall, 2008. We have awarded a fellowship to an incoming student who shows great promise for contributing to the field.

3. Internship Opportunity Development

Internship opportunities for our students are under development at several institutions, including the Smithsonian Institution, Purdue University Libraries, and the National Library of Medicine. The GSLIS Bioinformatics program is benefiting from our intern placement efforts: As there are not yet any DC students prepared for fieldwork opportunities, we have placed a bioinformatics student in a summer-long curation internship sponsored by the Johns Hopkins University Libraries and Computer Science Department. As we will have DC students preparing for graduation in the next academic year, we fully anticipate matching them with curation internship opportunities currently being established.

4. Needs Assessment

We modified our work on the first Needs Assessment to survey the Faculty of the Environment at the University of Illinois. We are conducting this investigation collaboratively with the Environmental Council (EC) at the University of Illinois at Urbana-Champaign. At this time, we have completed the survey instrument pre-test, and the Pilot Study was distributed the end of 2007. Based on the response and on interviews, we found that we needed to modify our sample for the final survey. We learned that there are many interesting variations in the kinds of data problems by discipline, and these were greater than anticipated. We reduced our list of survey respondents to include only those who *generate*

environmental data (above the molecular scale), giving us a total of 110 people, to whom we sent the survey in April 2008. We are currently awaiting the results, and expect follow-up interviews to help us analyze and aggregate responses by discipline. This data will, in turn, be used to facilitate our curriculum development and student training for supporting scientists in data management and curation activities.

5. Advisory Committee Activities

On March 28, 2008 we held our third Advisory Committee Meeting at the Missouri Botanical Garden in St. Louis, MO. Prior to the meeting we invited two new individuals to join the Committee in order to expand our local knowledge base, as well as to facilitate recruiting and internship opportunities.

The Advisory Committee members currently include:

- ◇ Chris Freeland, Application Development Manager and Project Manager, Missouri Botanical Garden
- ◇ Thomas Garnett, Associate Director for Digital Library and Information Systems, Smithsonian Institution
- ◇ Martin Kalfatovic, Head of the New Media Office and Preservation Services Department, Smithsonian Institution Libraries, Washington, D.C.
- ◇ Joanna McCaffrey, Collections Database Architect, The Field Museum
- ◇ Nancy McGovern, Digital Preservation Office, Inter-University Consortium for Political and Social Research (ICPSR), University of Michigan
- ◇ Katherine McNeill-Harman, Data Services and Economics Librarian, Massachusetts Institute of Technology
- ◇ Chuck Miller, Vice President, Information Technology and Chief Information Officer, Missouri Botanical Garden, St. Louis, MO
- ◇ David Soller, Ph.D., Geologist, U.S. Geological Survey, Reston, VA

At this meeting we presented to the Advisors the activities carried out and progress made on the grant project, including ongoing curriculum development, recruitment and internships, and the potential for a publishing opportunity for an edited monograph on data curation. During the afternoon session, the advisors worked with the project staff on framing two case studies that will be used in the classroom; one case is focused on biodiversity data collections, and the other on the development of the Data Documentation Initiative (DDI) 3 xml metadata specification for social science data. With the leadership of the Advisory Committee, we will be developing these case studies over the coming year, and plan to have them ready for use in the spring of 2009.

6. 2008 Data Curation Summer Institute

In our October Interim Report we described the ongoing requests that practicing librarians have made to DCEP faculty and staff for training or classes in data curation. To address this need, we will be conducting a 4-day Summer Institute on Data Curation, June 2-5, on the UIUC campus. This institute is being offered to 24-26 academic and research librarians who are involved in active work with research data in their university library.

The program for the week includes the following speakers and sessions:

- ◇ Introduction to Digital Data - John MacMullen, Graduate School of Library and Information Science (GSLIS), UIUC
- ◇ Selection and Appraisal - Melissa Cragin, GSLIS, UIUC
- ◇ Disciplinary Differences and "Talking with Domain Experts" - Carole Palmer, GSLIS, UIUC
- ◇ Integrity and Validation - Allen Renear, GSLIS, UIUC
- ◇ Preparing Data for Ingest - Ruth Duerr, National Snow and Ice Data Center

- ◇ Preservation Activities in Day-to-Day practice -Tim Donohue, UIUC Libraries IDEALS Repository
- ◇ Digital Preservation and Standards - Jerome McDonough, GSLIS
- ◇ Technical Aspects of Repository Systems for Data - Tim DiLauro, JHU Libraries
- ◇ Resource Requirements for Library DC Program - D. Scott Brandt, Purdue University Libraries
- ◇ Panel - Librarians and Scientists Working Together

7. Coordination and Dissemination Activities

The website was redesigned, and is located at <http://www.uiuc.edu/goto/dcep>. 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).

Ellen Rubenstein, the current Graduate Assistant (GA), worked on most of the activities described above, with particular focus on recruitment, the Needs Assessment, and planning for the Summer Institute for Data Curation, which is scheduled for June 2-5, 2008. Ellen will end her work on the project in July, and we will hire a new GA for the coming academic year.

The following list of presentations represent those that occurred during this reporting period:

Invited talks

Cragin, Melissa H. "Data Curation Services in Research Libraries." National Library of Medicine, April 15, 2008.

Panels

Cragin, Melissa H.; D'Avolio, Leonard; MacMullen, W. John; Smith, Catherine Arnott (2007). The Effects of Context on Data Quality in Biomedical Data Reuse. (Cragin's presentation: Data and Fitness for Re-use; MacMullen's presentation: Gene Ontology Annotations as an Example of the Impact of Curatorial variation on data reuse.) Panel at the 2007 Annual Meeting of the American Society for Information Science & Technology (ASIS&T).

Heidorn, P.B. Biological information management from molecules to ecosystems. Panel on "Biological Information Specialist Training," at 2nd International BioCurator Meeting, San Jose, CA, 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.

Posters

MacMullen, W. John (2007). Measuring variation in curators' GO annotations through a controlled multi-MOD study. Poster for the Second International Biocurator Meeting, San Jose, CA, October 2007.

MacMullen, W. John (2007). A Research Design for Measuring Variation in Database Curators' Annotations Through Prospective Randomized Controlled Studies. Poster for the 3rd International Digital Curation Conference, Washington, D.C. December 2007.

MacMullen, W. John. Understanding biocurators: Attributes and roles of model organism database curators. Poster for the Medical Library Association (MLA) Annual Meeting, May 2008.

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

Heidorn, P. Bryan, Palmer, Carole L., Wright, Dan & Cragin, Melissa. Biological Information Specialist's Training. Poster for the 2nd International BioCurator 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 for the Plant Science and Botany 2007 Conference, July 7-11, 2007, Chicago.

Other Activities

P. Bryan Heidorn, Ph.D. (PI) is now Program Director, Division of Biological Infrastructure, National Science Foundation. He serves on the e-Biosphere Steering Committee, and also on the Board of Directors of the JRS Biodiversity Foundation.

Melissa Cragin served on the Program Committee for the Third International Digital Curation Conference, Washington, D.C., December 11-13, 2007. Melissa has been invited to serve on the Program Committee for the Fourth Int'l. Digital Curation Conference, to be held in Edinburgh, Scotland, Dec. 1-3, 2008.

Carole L. Palmer, Ph.D., Associate Professor at the Graduate School of Library and Information Science, UIUC, attended 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

Related Work

In collaboration with the Purdue University Libraries, we developed an IMLS National Leadership grant proposal titled, "Investigating Data Curation Profiles across Research Disciplines." This application was successful, and we have begun 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.

We have been approached by the University of Illinois Press about developing a book on data curation, and we will explore this opportunity over the next year.