



An Educational Program on Data Curation

Melissa H. Cragin, P. Bryan Heidorn, Carole L. Palmer, Linda C. Smith

Graduate School of Library and Information Science

University of Illinois at Urbana-Champaign

Abstract

Several models of service are emerging in academic and research libraries for the collection and management of scientific data – an important segment of an institution’s total scholarly production. As libraries work to develop services to support the management of locally-generated data, they will require new kinds of expertise for providing appraisal, management, and access to data for long term use. Data curation is the active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education; curation activities enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time. At the Graduate School of Library and Information Science (GSLIS) at the University of Illinois at Urbana-Champaign our data curation curriculum will provide students with the theory and skills to plan and manage data curation systems, create and maintain data collections, and to evaluate and apply data and metadata standards for varied uses across the sciences, humanities, and social sciences.

Program Summary

The Data Curation Education Program (DCEP) is designed as a concentration within the ALA-accredited Master of Science (M.S.) degree at the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign. The M.S. requires a total of 40 credit hours of course work that includes required core courses. In addition to the courses required for all master’s students, those in the data curation concentration have two additional required core courses: **Foundations of Data Curation** and **Digital Preservation**. Courses recommended as electives include: Biodiversity and Ecoinformatics; Metadata in Theory and Practice; Museum Informatics; Ontologies (Humanities or Natural Sciences); Design of Digitally Mediated Information Services.

A major component of the DCEP will be participation in a **practicum** or **internship** related to the management, maintenance and preservation of data. We are developing a range of field experience sites in information-oriented institutions, including museums, data centers, libraries and institutional repositories, archives, and private industry.

“...creative partnerships between librarians and scientists will be critical for professional data stewardship in future science and engineering efforts.”

ARL (2006)

Anticipated Outcomes

- ♦ Identify best practices
- ♦ Create new courses and educational materials
- ♦ Develop a pool of field experience sites
- ♦ Cultivate new collaborative research partners
- ♦ Distribute curriculum materials via IDEALS, the UIUC institutional repository (e.g. syllabi and case studies)






Advisory Committee Contributions:

- ♦ Identify information problems and collect **best practices** from our partners to provide a broad understanding of information and data techniques, issues, and needs
- ♦ Facilitate **field placement** opportunities
- ♦ Develop **case studies** for use across the curriculum
- ♦ **Cultivate new partners** and new collaborative research

Cooperating Institutions

- ♦ Biomedical Informatics Research Network (UCSD)
- ♦ Marine Biological Laboratory
- ♦ Smithsonian Institution
- ♦ American Museum of Natural History
- ♦ Missouri Botanical Garden
- ♦ U.S. Army Strategic Environmental Management Program (SEMP)
- ♦ MIT Library

Related GSLIS Research

Automated metadata extraction and inference and Terminology, schema, and ontology development	Information roles and data management	Metadata and collections	Collaborative data collection	Preservation
				
Automatic Museum Label Metadata Extraction Heidorn, NSF DBI-9982849, NSF, DBI-0345387 Georeferencing Museum Specimen Sources Heidorn, Moore Foundation 2005-2929-00 Plant Description Standards Heidorn, IMLS NR-00-01-0017-01	Information and Discovery in Neuroscience Palmer, NSF IIS-0222848	Digital Collections and Content Cole, IMLS National Leadership Grant IG-02-02-0281	BioDiversity Survey Collaboration and Verification Heidorn and Palmer, NSF BDI-0113918	EchoDep (NDIPP) Unsworth, Sandore, Library of Congress
HERBIS automates herbarium specimen label imaging and data capture using OCR and machine learning. The process culminates with the population of label data and a specimen image into a structured collection database. The BioGeomancer Project is a worldwide collaboration of natural history and geospatial data experts developing automation to assist in georeferencing. Georeferencing is the process of converting text descriptions of locations to computer-readable geographic locations, such as a GIS system uses.	This project investigated the kinds of information work involved in scientific problem-solving. Through field studies at neuroscience labs we are identifying high impact information, critical information problems, and constraints on the transfer and exchange of information within research teams and between specializations and disciplines.	IMLS Digital Collections and Content (DCC) is a three-year effort at the University of Illinois to build an infrastructure for adaptable, interoperable, and sustainable digital collections.	TeleNature is a project designed to bring together scientists and citizen scientists for research and information sharing. The project is built upon wireless technology and individual enthusiasm for natural areas.	The ECHO DEpositary is a 3-year digital preservation research and development project at the University of Illinois at Urbana-Champaign in partnership with OCLC and funded by the Library of Congress under their National Digital Information Infrastructure Preservation Program (NDIIPP).

Future Events

- **Summer Institute in Data Curation** June 2-6, 2008 - workshop for academic and research librarians -
- International conference on the Practice, Problems and Promise of Data Curation in 2010, in collaboration with the UK Digital Curation Centre.

Association of Research Libraries (2006). To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering. A Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe. Washington, DC: ARL. Available at

The background image was created by Leo Reynolds under a Creative Commons 2.0 license (Attribution-NonCommercial-ShareAlike 2.0)

