Data Curation Education and Biological Information Specialists

P. Bryan Heidorn, Carole L. Palmer, Melissa H. Cragin, Linda C. Smith
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel St. Champaign, IL 61820
{pheidorn, clpalmer, cragin, lcsmith @uiuc.edu}

Abstract

Scientific data problems do not stand in isolation but are part of a larger set of challenges associated with the expansion of scientific information gathering capacity and changes in scholarly communication in the digital environment. These problems require new kinds of expertise in key areas, such as ontology development, data federation, and data visualization. However, recent reports on cyberinfrastructure and e-science initiatives acknowledge a shortage of qualified professionals to manage the increasing stores of data across the sciences (NSB, 2005). To build this kind of professional capacity, we have developed two complementary educational programs at the Graduate School of Library and Information Science (GSLIS) at the University of Illinois at Urbana-Champaign. One is the Biological Information Specialist (BIS) Master of Science degree, and the other is a concentration in Data Curation within the Master of Science in Library and Information Science. In this paper we discuss the key features of our Data Curation Education Program (DCEP), our approach to curriculum development, and the data curation program's contribution to the BIS curriculum. To provide a foundation for course development, we are identifying best practices in data curation, drawing from a variety of resources, including 1) a research foundation of information science projects in the biological sciences related to data curation; (2) a base of domain scientist collaborators; (3) active participation in disciplinary international standards development; (4) an advisory group representing bioinformatics, bench and field biosciences, and the information professions.

Curation for Integrative Biological Sciences

Scientific data problems are an integral part of the radical changes taking place in the practice of science. New scientific questions, new digital instrumentation, and new levels of interdisciplinary integration are leading to a dramatic increase in both data and derived information. Simultaneously, advances in communication options are leading to a transformation in scholarly communication in the digital environment. For the biological sciences, these complexities are increased because of the need to integrate data across scales. Many biologists are attempting to integrate or at least communicate their findings across scales of size, from molecules to organisms, and across ecosystems. In terms of time scales, some are working to integrate data from sub second molecular interactions through evolutionary time (see, for example, Wooley & Lin, 2005).

This work will require new kinds of expertise in key areas, such as ontology development, data federation, and data visualization. However, recent reports on e-research, cyberinfrastructure, and the stewardship of digital assets acknowledge a significant deficit in the workforce required to manage these burgeoning data stores. To address this growing need, with support from the Institute of Museum and Library Services (IMLS), the Graduate School of Library and Information Science (GSLIS) at the University of Illinois at Urbana-Champaign has initiated a data curation concentration within our ALA-accredited Master of Science degree. The Data

Curation Education Program (DCEP) is developing a specialized curriculum that focuses on data collection and management, knowledge representation, digital preservation and archiving, data standards, and policy. Our conception of data curation is consistent with Rusbridge, et. al. (2005), who state that digital curation includes not only data archiving and digital preservation but also active management and appraisal of data over the life-cycle of scientific interest. Students will be educated to take responsibility for assimilation and management of data in ways that add value and promote sharing across institutions and disciplinary specializations.

Effective data curation is a requisite component of multi-scale data integration and reuse. Therefore, in addition to being an elective concentration within our long-standing Master of Science in Library and Information Science, data curation is also foundational to our new Biological Information Specialist (BIS) program, which is part of a campus-wide Masters in Bioinformatics. In developing the BIS program, we consider our domain to be biological informatics, a broader field than bioinformatics. Biological informatics is concerned with the application of techniques to organize and integrate biological and abiotic information to solve multi-scale problems in biology. This may include, for example, integration of molecular bioinformatics with biodiversity informatics, geographic information systems, systematics, and atmospheric informatics to investigate problems such as the impact of genetic manipulation of crops on beneficial insect populations and long term crop yield (for example, the carabid beetle – Bt corn interaction, Lundgren & Wiedenmann, 2005). To develop professionals equipped to capture, maintain, and add value to biological data with long-term relevance at all scales, students in the bioinformatics masters program will obtain disciplinary training in biology subfields, giving them inside understanding of the needs of biologists, in addition to training in curatorial practice.

The concentration in data curation within the existing masters degree is different than the BIS program and does not require a background in science. The DCEP masters degree requires 40 graduate credits. Two courses are required for all masters students, and there are two additional required courses for the concentration, Foundations in Data Curation and Digital Preservation. DCEP students then select 2-4 electives from a designated set of courses that cover several areas such as digital libraries, knowledge representation and modeling, and informatics. Participating in a practicum experience is a recommended part of the program, and there will be opportunities for internships, as well. The section below presents the perspectives underpinning our approach to program development.

DCEP Program Features

Our data curation curriculum is based on three tiers of knowledge: knowledge of practice, knowledge of problems, and knowledge of promise. Current “practice” in data management has evolved to address emergent needs, thus there has been uneven development of curation practices across fields of scholarship. To extend the existing base of knowledge, best practices need to be identified and then assessed for application and adaptation to other fields of study. Additionally, many data “problems” have not yet been resolved by current practice, such as term disambiguation across neuroscience information systems, or the relationship between data set characteristics and levels or kinds of data curation required. Research to enumerate and analyze these problems can inform us about unmet needs in biology and other fields. Finally, the “promise” of new approaches to data curation to support future analytical purposes can be met through ongoing investigation of the interlinking of societal information trends and ever

changing technology.

Research Foundation
Our understanding of the range and variety of data curation roles that need to be addressed in our programs has been informed by our ongoing research activities in information technology and digital library development in the biological sciences. In recent years, we have worked closely with biological scientists, either collaboratively in technology development or cooperatively to learn more about information requirements, in field-based data gathering (Heidorn, Mehra, & Lokhaiser, 2002), digital flora and fauna (Heidorn, 2004; Greenberg, Heidorn & Seiberling, 2005), and neuroscience (Palmer, Cragin & Hogan, 2004, 2007). During that time, we have observed and documented information expertise that could have supported and helped advance how scientific research teams work with their data. Our research provides real-world cases and specific instances of data and information management problems that inform and function as teaching mechanisms for our curriculum. Other related GSLIS research projects also contribute in this way, including ECHODep, part of the Library of Congress National Digital Information Infrastructure Preservation Program (NDIIPP) (http://www.ndiipp.uiuc.edu/) and Digital Collections and Content, an IMLS funded digital collection registry and metadata repository. A number of these research projects have led to partnerships with individual practicing scientists and their institutions.

Collaboration with Domain Scientists / Scholars
An overarching objective of our educational initiative is the integration of research and practice through continued and new collaborations with scientists in various kinds of scientific institutions. As demonstrated by the research foundation outlined above, our approach to BIS and DCEP program development is highly dependent on the relationships we have developed with our research collaborators in the biological sciences. Current collaborating institutions include the Smithsonian Institution, the Missouri Botanical Garden, the American Museum of Natural History, the Marine Biological Laboratory (MBL), the Psychiatric Institute at the University of Illinois at Chicago, the Army Corp of Engineers, Engineering Research and Development Center, Construction Engineering Center, and the Biomedical Informatics Research Network (BIRN) at University of California at San Diego. These organizations are now serving as partners in our educational efforts.

Participation in International Standards Bodies and Data Sharing Projects
Both data curation and biological informatics are new fields of study. At this stage of development in these disciplines there is a natural progression from pilot projects to international standards. As professionals, we expect our students will be involved in standards initiatives, and therefore it is important that they receive training that is informed by current activities. To this end we have solicited input from the standards communities to aid in curriculum development and provide relevant experience for students. For example, for the past several years, Dr. Heidorn has been active with Biodiversity Information Standards: Taxonomic Database Working Group (TDWG) (http://www.tdwg.org) and the Global Biodiversity Information Facility (GBIF) (http://www.gbif.org). Students in the BIS program participated in the TDWG annual meeting through Internet audio and images. We worked with the GBIF education committee to help establish training objectives for both GBIF and TDWG and to identify skills that are critical to BIS and DCEP students. As an extension of this work, one student is currently working for GBIF on standards documentation. Finally, two of our faculty who are teaching courses for the DCEP participate on standards bodies that are fundamental to data curation: the

Metadata Encoding and Transmission Standard (METS; http://www.loc.gov/standards/mets/), and the FRBR - CIDOC CRM Harmonization working group (http://cidoc.ics.forth.gr/index.html).

Internships at National and Regional Data Centers
Internships will provide students with the opportunity to work directly with scientists and informatics experts on immediate, practical data and information problems. IMLS has provided funds to support internships at collaborating institutions, a number of which already run complementary internship programs. These "hands-on" experiences will be managed with both site supervision by practicing professionals as well as academic supervision from GSLIS faculty.

Advisory Committee Contributions
The first annual meeting of the DCEP Advisory Committee was convened in February 2007. The overarching goal of the advisory 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. 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.

To prepare for the meeting, advisory group members were asked to read Rusbridge, et al. (2005) and Palmer, et al. (2006), and review several additional resources, including the Digital Curation Centre's (DCC) curation manual (http://www.dcc.ac.uk/resource/curation-manual/chapters/). More importantly, each participant documented an actual data problem case from their institution for presentation during the meeting. We suggested that they consider in advance how their data problem related to the themes presented in the DCC curation manual. Discussion of the cases was semi-structured, with each presentation followed by commentary by fellow advisors and then by a request to enumerate the skill set required to meet the needs of the case study. After all advisors had presented their cases, the group addressed a series of questions and challenges related to the new curriculum. These included the potential role of practitioners and scientists in education, ranging from guest lectures to course teaching. In turn we solicited ideas for student recruitment, internships, job placement strategies, and approaches to widening best practice input, such as additional advisors and systematic data collection, and possible future research projects.

Extensive meeting minutes were recorded, which revealed a number of prominent data curation themes among the cases and related discussion. In addition, participants identified core skills for data curators during the course of the meeting. Centralized and distributed options for the location of data curation professionals were also considered. While one advisor advocated libraries as the best site for data curators, others expressed a need for data curators to be trained to work inside a variety of other research environments. Our program will be designed to

train professionals for both models, since we expect them to co-exist and be complementary in their implementation. Below we provide an overview of the primary themes that emerged from the presentations and exchange among the advisory group.

Practices and Problems in Data Curation

The knowledge and roles identified at our Advisory Committee meeting are presented in Table 1. They are organized by the themes that emerged from the discussion, and then differentiated by either technical or social orientation.

Table 1. Professional knowledge areas identified by the Advisory Committee.

Theme	Technical	Social
Data-centric	- File formats - Metadata - Ontologies - Standards	- Data practices
Data ownership	- Ownership models - Copyright variations	- Intellectual property issues
Policy and sustainability	- Preservation strategies	- Risk assessment
Researchers and research	- Workflows (and automation)	- Domain knowledge - Social parameters of research production - Research lifecycle - Scholarly communication
Management	- Use of applications across platforms (PC/Mac/Linux) - Basic data file care: storage, backup - Database management systems - Systems evaluation	- Personnel recruitment - Collaboration facilitation - Change management
Writing	- Grant applications - Scientific papers	- Grant proposal development

The items in the table represent the areas emphasized by the advisory group. Some areas were more prominent than others, and several items came up in discussion more than once. Moreover, this is not an exhaustive list, and the table does not account for variations in the nature of scientific data or disciplinary differences that affect data practices. Important distinctions exist that reflect technical and social domain-based requirements, practices, and cultures. Universal solutions will not meet every need, and the application of knowledge and skills to solve data curation problems will depend on institutional, legal, and policy parameters, as well as the data and its uses. We will continue to develop and refine this set of knowledge

areas, with the goal of determining a set of core skills.

In addition to professional knowledge, 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 practica and internships, the advisory group was eager to host students and expected that these placements could be mutually beneficial. In the course of this discussion, an interesting distinction was made between the functions of practicum versus internship placement: A practicum experience "should give the nuts and bolts of an institution," whereas internships should "take this to the next level, to contribute to research" or some kind of project.

Conclusions and Future Work

Digital data curation is a relatively new practice that must be aligned with the rapidly changing methods of science and scientific publishing. The data curation community must collaborate closely with scientists to identify their needs and opportunities for advancing scientific practice. In our engagement with the scientific community thus far, we have gained insights into the "practice" and "problem" tiers of data curation knowledge. This will help us to establish the education requirements for data curation and the basic foundation for educating a new generation of professionals to serve the scientific enterprise. But, this does not take us far enough. To fully exploit the growing capital of high-quality, curated data, we must also develop programs of research to further investigate the "promise" of data curation for supporting integrative and data-driven research.

Acknowledgements

This project has been supported by the following grants: Centuries of Knowledge, IMLS RE-05-05-0036 (P.B. Heidorn, PI) and A Graduate Program for Scientific Communication Specialists: Getting Past the Prototype in Biological Informatics, NSF-IIS-0534567 (C.L. Palmer, PI). We thank Ellen Rubenstein for assisting with preparation and analysis of the Advisory Committee meeting transcript.

References

Greenberg, J., Heidorn, P. B., & Seiberling, S. (2005). Growing Vocabularies for Plant Identification and Scientific Learning. International Conference on Dublin Core and Metadata Applications (DC-2005, Sept 15, 2005), Madrid, Spain.

Heidorn, P. B. (2004). Publishing Digital Floras and Faunas. Bulletin of the American Society for Information Science & Technology, 30 (2), 8-11.

Heidorn, P. B., Mehra, B., Lokhaiser, M. (2002). Complementary User-Centered Methodologies for Information Seeking and Use: System's Design in the Biological Information Browsing Environment (BIBE). Journal of the American Society for Information Science & Technology, 53(14), 1251-1258.

Lundgren, J. G., Wiedenmann, R. N. (2005). Tritrophic Interactions among Bt (Cry3Bb1) Corn, Aphid Prey, and the Predator Coleomegilla maculata (Coleoptera: Coccinellidae). Environmental Entomology, 34(6), 1621–1625.

National Science Board. (2005). NSB-05-40, Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century. http://www.nsf.gov/pubs/2005/nsb0540/

Palmer, C. L., Cragin, M. H., and Hogan, T. P. (2004). Information at the Intersections of Discovery: Case Studies in Neuroscience. Proceedings of the American Society for Information Science & Technology annual meeting, 41, 448-455.

Palmer, C. L., Cragin, M. H., & Hogan, T. P. (2007). Weak Information Work in Scientific Discovery. Information Processing & Management, 43(3), 808-820.

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

Rusbridge, C., Burnhill, P., Ross, S., Buneman, P., Giaretta, D., Lyon, L., & Atkinson, M. (2005). The Digital Curation Centre: A Vision for Digital Curation. In Proceedings of Local to Global Data Interoperability - Challenges and Technologies. Mass Storage and Systems Technology Committee of the IEEE Computer Society, June 20-24, Sardinia, Italy. http://eprints.erpanet.org/82/

Wooley, J. C., & Lin, H. (2005). Catalyzing Inquiry at the Interface of Computing and Biology. Washington, D.C.: National Academies Press.