The International Journal of Digital Curation
Issue 2, Volume 2 | 2007

Graduate Curriculum for Biological Information Specialists: A Key to Integration of Scale in Biology

Carole L. Palmer, P. Bryan Heidorn, Dan Wright, and Melissa H. Cragin,
Graduate School of Library and Information Science,
University of Illinois at Urbana-Champaign

November 2007

Abstract

Scientific data problems do not stand in isolation. They are part of a larger set of challenges associated with the escalation of scientific information and changes in scholarly communication in the digital environment. Biologists in particular are generating enormous sets of data at a high rate, and new discoveries in the biological sciences will increasingly depend on the integration of data across multiple scales. This work will require new kinds of information expertise in key areas. To build this professional capacity we have developed two complementary educational programs: a Biological Information Specialist (BIS) masters degree and a concentration in Data Curation (DC). We believe that BISs will be central in the development of cyberinfrastructure and information services needed to facilitate interdisciplinary and multi-scale science. Here we present three sample cases from our current research projects to illustrate areas in which we expect information specialists to make important contributions to biological research practice.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

Recent reports on cyberinfrastructure and e-science initiatives recognize a shortage of qualified professionals to manage the increasing stores of data across the sciences (National Science Board [NSB], 2005). Assessments in the UK point specifically to the critical shortage of trained personnel to carry out digital curation activities and to a lack of training programs to supply those personnel (Lord & Macdonald, 2003;Cornwell Management Consultants plc., 2004). But scientific data problems do not stand in isolation. They are part of a larger set of challenges associated with escalated production of scientific information and changes in scholarly communication in the digital environment. In the biological sciences, ranging from protein structure prediction to neuroscience to biodiversity, researchers are producing and consuming increasing amounts and varieties of information and data, while striving to work with these resources in new ways. This has lead to daunting problems with information management and integration.

Biology has become an extremely data-intensive science, and there are numerous challenges associated with the amount and rate of data being generated. However, the complexity of the underlying biology reflected in the burgeoning body of data is of greater consequence for scientific discovery than the volume. It has been recognized that the future success of the field lies in an integrative approach to solving biological problems (Woese, 2004; Wooley & Lin, 2005). Moreover, new discoveries in the biological sciences will increasingly depend on the integration of data across multiple scales – of size, time, and orders of complexity. Researchers will draw on data from other disciplines to gain new insights into their own research questions. One example of this kind of work is in systems biology. An integrated picture of all the processes in a cell requires data ranging from sub-atomic to microscopic scales from a number of different domains, and this range of data must be acquired from a variety of sources (Wooley & Lin, 2005). To enable this cross-scale, interdisciplinary integration for the coming generations of biological researchers, data must be managed to facilitate interoperability, preservation, and sharing. Best practices and data standards already exist or are in active development, but a majority of scientists are unaware of (and sometimes uninterested in) issues such as metadata formats and interoperability.

This situation calls for a new breed of information professional trained in best practices of biological information collection and management, and who is knowledgeable about the differences and commonalities of these practices across domains and can promote interoperability and sharing. 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, a Biological Information Specialist (BIS) masters degree and a concentration in Data Curation (DC). Our approach to BIS and DC education is grounded in the recognition that while the volume of information is escalating in the digital environment, the character of information and research is also changing.

This situation calls for a new breed of information professional trained in best practices of biological information collection and management, and who is knowledgeable about the differences and commonalities of these practices across domains and can promote interoperability and sharing. 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, a Biological Information Specialist (BIS) masters degree and a concentration in Data Curation (DC). Our approach to BIS and DC education is grounded in the recognition that while the volume of information is escalating in the digital environment, the character of information and research is also changing.

The degree is offered as part of a new campus-wide Masters in Bioinformatics initiative[1], but the GSLIS scope is very different from other bioinformatics programs being developed on our campus and elsewhere. Most bioinformatics programs focus on computational molecular biology, however bioinformatics has been broadly construed in segments of the scientific community as applying to all scales of biological data, as evidenced in the National Institute of Health, Biomedical Information Science and Technology Initiative (BISTI) definition of bioinformatics:

> Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data. (Biomedical Information Science and Technology Initiative Consortium [BISTIC] Definition Committee, 2000)

Existing bioinformatics programs have tended to concentrate on the analysis and visualization areas. Our program covers all areas outlined in the BISTIC definition, including a focus on expanding use, and re-use, of information and data.

Curation for Integrative Bioscience

While our data curation concentration (DC) will be an option in our long-standing Masters of Science in Library and Information Science program, it is also a foundational component of the BIS program, since effective data curation is a prerequisite for multi-scale data integration and re-use. 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. "This adds value through the provision of context and linkage, placing emphasis on publishing data in ways that ease re-use and promoting accountability and integration." Students with the DC concentration will be educated to take responsibility for assimilation and management of data in ways that add value *and* promote sharing across laboratories and disciplinary specializations. This broad focus is necessary for capturing and maintaining the long-term relevance of biological data at all scales.

In some disciplines, such as astronomy and high-energy particle physics, work on data management and curatorial problems has been an integral part of the scientific endeavor for a number of years. In other domains, however, the concepts and methods that constitute data curation are still new and will require articulation and integration into daily practice. In the life sciences, for example, curation concerns have been addressed in a segmented fashion. Standards work has been carried out by various specialized groups in a "bottom-up" fashion, and querying techniques are largely

[1] Please see http://sci.lis.uiuc.edu/ for more information on the GSLIS and UIUC campus initiatives.

addressed by database, data mining, and statistics experts (Olken & Jagadish, 2003). Preservation is often equated with the creation of an "archival back-up," and concepts such as provenance, presentation for re-use, and workflow capture are rarely addressed. While the database community controls many of the data modeling and database design tasks, other curatorial work falls outside of their role. For example, according to Jagadish & Olken (2004):

> …it has become increasingly clear that good data management infrastructure for recording and querying data provenance – the origin and processing history of data – is vital if we are to effectively encourage the sharing of biological and biomedical data. Data provenance issues have been largely neglected by the database research community except for a few researchers in statistical data management and data warehousing (Jagadish & Olken, 2004, p. 18).

While it is possible that minor adaptations will suffice to meet some of these challenges in the genomics field (Markowitz, 2003), it is unlikely that DBMS (database management system) approaches will meet the challenges posed throughout the data life-cycle for the range biological sciences. Expertise is needed in additional key areas, including scientific metadata, ontology and standards development, and interoperability.

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 these and previous projects, we have worked closely with biological scientists, either collaboratively in technology development or cooperatively to learn more about information requirements. During that time, we have observed and documented the information expertise that could have supported and helped advance how scientific research teams work with their data. Other GSLIS research projects are also feeding into our base of knowledge and expertise, especially the ECHODep digital preservation project, funded by the Library of Congress under their National Digital Information Infrastructure Preservation Program (NDIIPP)[2] Our research provides real-world cases and specific instances of data and information management problems around which to develop our curriculum. Here we present three sample cases that illustrate areas in which we expect information professionals to contribute to biological research practice.

Modeling and Computational Neuroscience

Modelers are a known user group for the extensive stores of data being generated by biologists. Experimental data are a necessity for modelers, yet they will rarely – if ever – generate their own data sets. In neuroscience, modelers are computational neuroscientists, mathematicians, even physicists (among others), who are interested in solving biological questions by applying computer algorithms to composite data sets to simulate behavior of components in biological systems, from subcellular (molecular) to the organismic level.

[2] http://www.ndiipp.uiuc.edu/

Modeling communities are not always directly connected with experimental scientists, and therefore may not know where to find workable data sources. Once data are located, it is generally the case that modeling work was not considered during the planning and collection of a given experimental data set, resulting in barriers to use for these communities. In addition, while the use of literature to build models is helpful – modelers can use descriptions of structures and functions of subcellular components to develop frameworks – literature sources are not enough. As described by one computational neuroscientist modeling how neurons communicate via synapses: "Based on the nature of simulations, there are more stringent requirements on the data. (Other) labs are doing reconstructions on a coarse level, and you can't use some reported" in past literature. Even when they have done serial reconstructions, "when you contact them or read closely," their model isn't water tight, "it has holes; they're interested in (more) coarse measurements. We would have to re-create a lot of work to use it."

Automated Metadata Extraction and Inference

In botany, entomology, zoology and other fields, historically significant collections of scientific objects have been curated for centuries along with their associated metadata. A clear example is museum specimens. We have been involved with two large collaborative projects with multiple institutions to extract and transform this information for both traditional and unplanned, future uses. In the HERBIS Project,[3] museum herbarium specimens are imaged and optical character recognition or hand writing recognition performed on the specimens labels. The text from this OCR is processed through machine-learning algorithms that have been trained to extract Darwin Core metadata and other information from the stream of text to produce an XML document codifying this information for verification and storage in databases. Data associated with the objects include changes in species names over time with multiple determinations, collection dates, locations and other essential descriptive information.

The Biogeomancer Project[4] converts natural language descriptions of locations into latitude and longitude and calculates uncertainty intervals. Spatial descriptions such as "Baird Mtns.; Salmon R. headwaters" are automatically converted to sexagesimal coordinates 67°45'21"N 159°29'46"W. Biogeomancer works on any locality data not just herbarium specimens. Once the work of HERBIS and Biogeomancer is performed the data can be put to new uses. For example, after sufficient data cleaning and validation (Chapman, 2005), maps can be automatically generated to depict the historical distribution of species over time to be compared to the current distribution. Products such as the developing TCS (Taxonomic Concept Schema) can help to map old names on labels with current usage. ABCD (Access to Biological Collection Data) and Darwin Core, along with the DigIR federation protocol, are used to make the data globally available through the Global Biodiversity Information Facility and other data clearinghouses. These larger collections can in turn be mapped to environmental models and geological surface conditions to predict ranges under climate change scenarios, an application 18th century scientists never would have dreamed of when they first carefully recorded information about a newly collected specimen.

[3] http://www.herbis.org/index.php
[4] http://www.biogeomancer.org/

3.3 Ontology Development

Current ontology development work is aimed at serving several goals, including knowledge representation, discovery, and data integration. In the field of biodiversity, the Taxonomic Databases Working Group (TDWG) is developing a Biodiversity Informatics Core Ontology.[5] While the ultimate goal is to span all of biodiversity, the initiative began with an effort to produce a unified ontology to encompass four biodiversity schema: ABCD, Darwin Core, SDD (Structure of Descriptive Data) and TCS. ABCD is an evolving comprehensive standard containing over 700 elements for access to and exchange of data about specimens and observations, or primary biodiversity data. The Darwin Core is a standard designed to facilitate the exchange of information about the time and geographic occurrence of species, consisting of only 40+ elements to simplify data interchange. SDD supplies a framework for description of biological entities of any type, primarily for interchange among interactive key systems and taxonomic description publishing. Under TCS, taxonomic names are part of a taxon concept and do not exist independently; the taxonomic concept captures the relationships and beliefs associated with the broader context of taxonomy. Computer scientists, biologists, museum directors, librarians and others are working together to manage the complexity of this undertaking.

Similar ontology efforts are underway in the neurosciences. Two prominent and practical objectives in this work are to integrate biological data across scales and to link animal and human imaging data. For example, researchers with the Biomedical Information Research Network (BIRN) are developing tools to integrate data and knowledge structures to facilitate these activities. The ontology under development will be used by both humans and machines, and needs to support user groups that include anatomists, neuroscientists and neurologists, pathologists, and genomics researchers. Considerable work is going into vocabulary control, which has proven to be of great importance for activities such as data annotation.

Since existing vocabularies have not been complete or accurate enough for outright application to the integrated data system, the BIRN community of biologists is working on its own, the BIRNLex. This work, in conjunction with the creation of the ontology to connect data and knowledge systems, has been highly labor-intensive for the participating scientists and associates, and, therefore, very costly. One meeting we attended, which was supported by video conferencing, included three PI level biologists, three research scientists, a number of graduate students, and a representative from a national funding agency who flew in to be on site.

Our programs will provide fundamental training in areas at the heart of these kinds of activities, including knowledge representation and organization, classification, data modeling, and ontology development. As a case in point, ontology education will aim to develop knowledge of the spectrum of biomedical ontologies to facilitate integration of data collections (databases) and resolve term and definitional conflicts. Our approach strives to balance the philosophical or theoretical ontology perspective and real world (semantic) use. In relation to data curation, the higher demands for functionality of data means that information professionals will need to be trained to contribute from the time of data creation through management and later re-use, because of the dynamics and the need for persistent relationships among data stores around the world.

[5] http://wiki.tdwg.org/twiki/bin/view/TAG/TDWGOntology

LIS Orientation

As a long-standing leader in LIS (Library and Information Science) research and education, GSLIS is well positioned to advance these new programs. LIS is the only field that is concerned with the full landscape of scientific information and the interactions therein, and with the provision of services to exploit that base of information (Bates, 1999; White, Bates, & Wilson, 1992). Moreover, LIS has a tradition of training information professionals to work in scientific research settings. Our conception of the BIS is an extension of the informationist movement that began within LIS over 30 years ago. Beginning with an emphasis on clinical medical librarianship, informationists have now advanced beyond the clinical realm to also work as members of scientific research groups toward similar goals of improving information use and communication among teams. Moreover, there has been a growing recognition that expertise in both information science and the research domain is crucial for information professionals contributing to scientific research (Florance, Bettinsoli, & Ketchell, 2002). In some cases the individuals entering the BIS program will have prior training in either biology or information science at the undergraduate or graduate level. In order to balance their skills as BISs, they will need to bridge both disciplines.

LIS also has a strong tradition of focusing on the information needs of users rather than on internal system criteria, such as the technical elegance of software. In this way, LIS-trained professionals are well prepared to collaborate with biological scientists. They accept that research practices of scientists must be understood and accommodated for effective design of information systems, and, in programs such as ours, are trained in empirical techniques for studying domain-based research practices. BISs will have appropriate training to solve information problems in concert with scientists and will complement, not duplicate, the expertise of computational scientists. While computer science is vital for advancing the state-of-the-art in computational biology, BISs will be central in developing the cyberinfrastructure and information services necessary to facilitate interdisciplinary and multi-scale science.

Conclusion

An overarching objective of our educational initiative is the integration of research and practice through continued and new collaborations with scientific partners. As demonstrated in the research cases presented above, our approach to BIS and DC 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, Missouri Botanical Garden, American Museum of Natural History, the Psychiatric Institute at the University of Illinois at Chicago, and BIRN at University of California at San Diego. These organizations are now serving as partners in our educational efforts. With the assistance of these advisors and others in the data curation community, we are collecting best practices for BIS and DC curriculum development, and will be cumulating and publishing them for the larger scientific and LIS communities.

Our advisory group provides a broad perspective on information and data practices and problems in the biological sciences that directly informs our curriculum. In addition, we are conducting interviews and surveys of scientists to further document needs and practices across a broader range of biological sciences. Additional technical

requirements have been determined through position descriptions from jobs focused on the management of scientific data. For example, job advertisements posted on the Taxacom and TDWG listservs provide details on the needs of the biological taxonomy community. The jobs in this field range from relatively stable federal government posts to two-year grant-funded project contracts with the hope of renewal. The skills identified through these various means coincide to a great degree and together represent a highly complementary array of professional biological data and information management skills. In addition to the more general requirements of technical communication skills and subject knowledge, we have found a demonstrated need for expertise in data archiving and preservation; management of instrumentation data; scientific databases, data repositories, and tools; data and metadata standards; biological ontologies; workflow capture; data synthesis; literature-based discovery; and copyright and intellectual property issues.

These content areas and other related topics are being addressed in a range of new and existing courses. Two new core courses have been developed specifically for the data curation specialization: Foundations of Data Curation and Digital Preservation. Other semester-long courses offered by the school include Metadata in Theory and Practice, and Biodiversity and Ecoinformatics.[6] Two new courses designed for the BIS program, Ontologies in the Natural Sciences and Introduction to Biological Informatics Problems and Resources, provide further coverage and add depth in their specified areas. While there may be universal principles of best practice that can be covered in coursework, data curation needs and practices vary broadly across the field of biology and all disciplines, making practical workplace training indispensable. Therefore, students will be able to focus their thesis work and participate in practica or internships in particular types of institutions where they can work more directly with scientists and informatics experts on practical and immediate data and information problems.

Through the training and placement of Biological Information Specialists in research institutions, we aim to facilitate and encourage widespread adoption of best practices for supporting the information and data needs of the biological sciences. Our activities should also cultivate new research projects with our partners that will continue to inform our understanding of the future roles of information professionals in the advancement of science.

Acknowledgements

This work was supported in part by grants from the Institute of Museum and Library Services RE-05-06-0036-06 and National Science Foundation IIS 0534567, IIS 0222848, and DBI 0345387. We also wish to thank our colleagues, Allen Renear, Dave Dubin, and others involved with the GSLIS research writing group, for their contributions during the development of this paper.

[6] The syllabus is available at https://netfiles.uiuc.edu/pheidorn/www/LIS590BDL2007/

References

Bates, M. J. (1999). The invisible substrate of information science. *Journal of the American Society for Information Science, 50(12)*, 1043-1050.

BISTIC (Biomedical Information Science and Technology Initiative Consortium) Definition Committee. (2000). NIH working definition of bioinformatics and computational biology. Retrieved December 4, 2007 from http://www.bisti.nih.gov/CompuBioDef.pdf

Chapman, A. D. (2005). *Principles and methods of data cleaning – primary species and species-occurrence data, version 1.0*. Report for the Global Biodiversity Information Facility, Copenhagen. Retrieved December 4, 2007 from http://www.gbif.org/prog/digit/data_quality/DataCleaning.pdf

Cornwell Management Consultants plc. (2004). *Digital preservation coalition training needs analysis final report*. A report funded by the Joint Information Systems Committee. Retrieved December 4, 2007 from http://www.jisc.ac.uk/uploaded_documents/finalReport.pdf

Florance, V., Bettinsoli, G., & Ketchell, D. S. (2002). Information in context: Integrating information specialists into practice settings. *Journal of the Medical Library Association 90(1)*, pp.49–58.

Jagadish, H. V., & Olken, F. (2004). Data engineering for life sciences: Database management for life sciences research. *SIGMOD Record, 33(2)*, 15-20.

Lord, P., & Macdonald, A. (2003). *Data curation for e-science in the UK: an audit to establish requirements for future curation and provision*. E-Science Curation Report. JISC Committee for the Support of Research. Retrieved December 4, 2007 from http://www.jisc.ac.uk/uploaded_documents/e-scienceReportFinal.pdf

Markowitz, V. M. (2003). Data management challenges for molecular and cell biology: An industry perspective. *OMICS, 7(1)*, 121-122.

National Science Board. (2005). *Long-lived digital data collections enabling research and education in the 21st century*. NSB-05-40. Retrieved December 4, 2007 from http://www.nsf.gov/pubs/2005/nsb0540/

Olken, F., & Jagadish, H. V. (2003). Data management for integrative biology. *OMICS, 7(1)*, 1.

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, 2005*. Mass Storage and Systems Technology Committee of the IEEE Computer Society, June 20-24, 2005, Sardinia, Italy. Retrieved December 4, 2007 from http://eprints.erpanet.org/82/

White, H. D., Bates, M. J., & Wilson, P. (1992). For information specialists: Interpretations of reference and bibliographic work. Norwood: Ablex.

Woese, C. R. (2004). A new biology for a new century. *Microbiology & Molecular Biology Reviews 68(2)*, 173-186.

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