

## **A Graduate Program for Scientific Communication Specialists: Getting Past the Prototype in Biological Informatics**

Carole L. Palmer and Bryan Heidorn  
Graduate School of Library and Information Science  
University of Illinois at Urbana-Champaign

National Science Foundation/IIS/CISE Education Research and Curriculum Development  
January 2006-January 2009

### **Project Summary**

We propose to develop a comprehensive masters level training program in scientific communication. The program will build on our long-standing, top-rated masters of science in library and information science (MSLIS), our new digital library certificate of advanced studies program (DL-CAS), and our ongoing information technology research initiatives in biodiversity, neuroscience, and genomics.

**Intellectual merit:** The nature of science communication is changing to include new forms of information collection, management, use, and dissemination, and biologists are generating new information at a staggering rate. Methods of information management, such as globally federated data sets in ecology and neuroscience not only alter the quantity of information that is available but also cause a qualitative shift in the nature of the scientific questions that can be asked and answered. For example, when information on the historic global distribution of species is combined with historic climatic data we can begin to answer questions about the future distribution of species under the pressure of climate change. This data transformation however is not free. It requires the collaborative effort of scientists with information specialists who both understand some aspects of the science and who are skilled in the associated computational and information management tasks. This new breed of information specialist must know how to collect, organize, access, use, and preserve information, freeing biological scientists to focus on biology.

Current bioinformatics programs are computation centric and focus primarily on molecular biology. While they make a valuable contribution to science, greater progress in research can be made by expanding the area of scientific and information technology expertise covered by information professionals supporting scientific research. The GSLIS program will be unique in that the main focus is on training scientific communication specialists (SCS) within a much broader scope of biological informatics to include integration of scale from the biomolecular to ecosystem. Ongoing biological information systems research will be combined with the graduate training program to give students hands on, real life experience with current scientific research initiatives.

**Broader Impacts:** Currently, information management tasks are performed either by biological scientists who are self taught in information management or by computer scientists with great computational training but limited understanding of the problem domain. SCSs will play an important role in improving information transfer and collaboration in science. They will allow biological scientists to concentrate on scientific problems and computer scientists to let go of projects when they move into implementation stages.

The project will build on ongoing local and national collaborations to support curriculum development and internships. Students, teachers and researchers working together will be able to develop new tools and methods for research and education, and these relationships will fuel new collaborative research opportunities. By working with preeminent partner institutions in biological informatics, we will be able to bring together and publish the best practices that have developed independently in the biology research projects around the country. The resulting curriculum will be published and made available on the web. Most importantly, the students will graduate and spread their training to research groups around the country.

## **Project Description**

### **1. Educational aims**

We propose to develop a comprehensive masters level training program in scientific communication. The program will build on the Graduate School of Library and Information Science (GSLIS) long-standing, top-rated master's of science in library and information science (MSLIS), our new digital library certificate of advanced study program (DL-CAS), and our ongoing information technology research initiatives in biodiversity, neuroscience, and genomics. It will be part of a new campus-level bioinformatics program at the University of Illinois at Urbana-Champaign (UIUC) that aims to provide students with the interdisciplinary skills needed for digital age science. The computer science department, chemical and biomedical engineering department, and college of agricultural, consumer, and environmental sciences are offering degrees that will prepare students for work in biomolecular informatics. The GSLIS program will be unique in that the main focus is on training scientific communication specialists (SCS) within a much broader scope of biology to include integration of scale from the biomolecular to ecosystems. NSF support will allow us to provide a more comprehensive program as well as more complete evaluation and national dissemination of the results.

Why do we see a need for a masters degree program focused on scientific communication and rooted in Library and Information Science (LIS)? Simply stated, scientists should be able to spend time conducting science. Of course, doing science often involves information technology development, and, as Kling & McKim (2000) demonstrated, information systems should grow out of the needs and cultures of research communities. Information technology should not be based on vague notions that conflate distinct activities and interests of different research domains. However, there are many parts of information technology development and sustainability, as well as other information-based activities in the daily practice of science, that could benefit greatly if supported by specialists devoted primarily to the information, in service to the science.

In the future of informatics development, SCSs will complement, not duplicate, the expertise of biological and computational scientists. They would serve as supportive team members to keep information technology development moving "past the prototype", and they would continue to play their traditional research librarian roles in advancing research and scholarly communication. There is an extensive and ever growing universe of information resources, informatics tools, and scholarly communication options that need to be understood, assessed, and coordinated locally so that there can be more rational and equitable global development in terms of integration of data, literature, and information technologies. SCSs will bridge these arenas of informatics development.

Bioinformatics programs are being developed across the country; however they do not cover in a comprehensive way the range of information issues that have emerged in the past decade. A few examples include data exchange standards, digital preservation, and electronic publishing. Moreover, they are not focused on the larger family of information problems that cut across all the sciences. While we understand the great need for discipline-specific specializations, it is important to recognize that the scientific enterprise will be more effective with well-trained professionals working to systematically advance information use at local and global levels.

#### **1.1 Goals and objectives**

Our primary goal is to design a program of graduate study that can serve as a model for training specialists in scientific communication for biological informatics. A secondary goal is to integrate this graduate training with ongoing bioinformatics research and practice to produce specialists that understand the research culture and can make substantive contributions to scientific discovery. The objectives are to:

- 1) develop curriculum for the specialization that builds logically on existing graduate programs at GSLIS and the core provided by the new campus-level bioinformatics degree;
- 2) establish internships at institutions and laboratories where students can develop and apply their growing expertise;
- 3) develop mechanisms for integrating course work into current informatics research at GSLIS, other UIUC departments, and partner institutions;
- 4) share the educational approach to biological informatics with other schools interested in developing similar specializations;
- 5) expand understanding of the role of informatics in scientific progress.

## 1.2 Broader impacts

We expect SCSs to have important local impacts on the scientific enterprise while also contributing to global development and integration of information systems for scientific research and communication. Their levels of responsibility will vary depending on their placement, in a particular laboratory, research institute, academic department, or unit of a research library. However, their influences may be wide-reaching.

Local impacts will include: 1) Progress toward what we call the “getting past the prototype” problem. SCSs will attend to implementation, evaluation, continual improvement, and sustainability of information and data systems. That is, they will make sure systems are responsive to the real needs of scientists and therefore assure greater adoption of technologies by target communities. 2) SCSs will be prepared to better exploit existing data standards and work toward “long-lived data” and interoperability across the various scientific communities. They will be instrumental in building and integrating the increasing number of ontologies, taxonomies, digital libraries, indexing systems, and vocabularies associated with digital data and products.

Global impacts will include: 1) More equitable access to publishing forums by scientists in general. Many of the larger, resource-rich research institutions are moving to an open access institutional repository model for disseminating research. Scientists at smaller institutions are likely to fall behind in this trend or be under-represented in such repositories. SCSs can play an important role on many campuses by developing and managing these new communication forums and assisting scientists in disseminating and preserving their intellectual property. They will stay up to date on and participate in data sharing and federation activities within and across scientific disciplines where preservation and archiving of data and results now requires ongoing communication and collaboration.

At both the local and global level, SCSs will play an important role in improving information transfer and collaboration in science. They will allow biological scientists to concentrate on scientific problems and computer scientists to let go of projects when they move into implementation stages.

## **2. Background**

### 2.1 Bioinformatics and biological informatics

What is bioinformatics? While there are many ways to characterize bioinformatics, we interpret the field broadly to include the range of biological sciences and a broad conceptualization of information. While bioinformatics is frequently associated with data mining and molecular modeling, all the biological sciences are moving forward with computational approaches, and they all are facing increasing problems related to finding, mobilizing, preserving, standardizing, sharing, and managing information and data. According to the NIH Biomedical Information Science and Technology Initiative (BISTI) documentation, bioinformatics is:

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.

This definition encompasses the range of biological sciences covered in our current science-oriented research projects in our college and the information problems that are at the core of our existing curriculum. It captures aspects of information production and use that are important in the entire cycle of scientific research and communication, from data collection to dissemination of research results.

The scope of this definition also includes biodiversity or biological diversity, a field often overlooked in discussions of bioinformatics. Biological diversity means “the variability among living organisms from all sources, including inter alia, terrestrial, marine and other aquatic ecosystems and the ecological complexes of which they are part; this includes diversity within species, between species and of ecosystems” (Convention on Biological Diversity, Art. 2, para. 1). Informatics is as vital to biodiversity biologists as it is to molecular biologists. As E. O. Wilson states, biologists are turning to information technology to produce critically needed efficiencies in their work but much more effort is needed.

New electronic technology, increasing exponentially in power, is trimming the cost and time required for taxonomic description and data analysis. It promises to speed traditional systematics by two orders of magnitude. What is lacking and needed now is a concerted effort, comparable to the Human Genome Project (HGP), to complete a global biodiversity survey – pole to pole, whales to bacteria, and in a reasonably short period of time (Wilson, E.O., 2000).

Unfortunately, most of the information science community is unfamiliar with these biological fields of study so definitions and explanations are in order. The relationship between “bioinformatics” and the term that we have chosen to use here, “biological informatics,” is not as subtle as it might seem. Over the past ten years, the term, “bioinformatics” has come to mean, “information on molecular biology” and in particular gene and protein sequences. This use of the term in the popular press, associated with the great progress and successes in that field has served to cement this definition into the psyches of the general population and scientists alike—thus the need for a new term (biological informatics) to cover the science of information about all levels of biological analysis. Health Informatics, medical informatics, neuroinformatics, biodiversity informatics, and biomolecular informatics all fall under this broader concept (Heidorn, 2003a).

Not only has information technology changed the face of biology on the local level for scientists, but the field has undergone a revolutionary globalization and shift in scale which has introduced new challenges for bioinformatics. The Convention on Biological Diversity is an example at the legislative level where countries attempt to coordinate their efforts. On the operational level institutions such as the Global Biodiversity Information Facility (GBIF) (<http://www.gbif.org>), the International Union of Biological Sciences Taxonomic Database Working Group (TDWG) ([www.tdwg.org](http://www.tdwg.org)) and many others work to develop international standards for data sharing. Previously it was sufficient for individual researchers in individual laboratories to store information in local formats. Because of the scale of science it is now desirable to share data globally. The burden of this shift can not fall to the shoulders of the biological scientists alone but should fall to a new breed of professionals versed in global biodiversity information demands.

## 2.2 Scientists and information technologies

Over the past several years, GSLIS has been building a research concentration in information technology and digital library development for domains in the biological sciences. Projects include Carole Palmer’s NSF IIS-0222848, Bryan Heidorn’s NSF DBI-9982849, NSF, BDI-0345387, IMLS NR-00-01-0017-01, Moore 2005-2929-00, Heidorn and Palmer’s, NSF BDI-0113918, and Bruce Schatz’s NSF IRI-90-15407, NSF IRI-92-57252, NSF BIR-93-19844, and

NSF FIBR 0425852.) 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 continually observed and documented cases where LIS expertise could have supported and helped advance the work of scientific research teams. Here we offer our perspectives based on our recent neuroscience, biodiversity, and museum informatics projects.

Collaborative neuroinformatics projects offer excellent examples of how scientists from a biological domain and computer scientists bring together essential and complementary expertise to develop innovative scientific technologies. However, as projects progress the focus necessarily shifts from work on the hard research problems that relate to the biological and computer science research domains to refining and building content and functionality of the system and promoting adoption. These later stages of work are fundamental to advancing scientific practice, but they are not central to the objectives of practicing biological or computational scientists. In our case studies of discovery processes in neuroscience, we have come to call this the “getting past the prototype” problem.

There is a tendency for collaborations to languish once the prototype is produced. System development does not progress well after this point and therefore technologies are not transformed into products that can be readily applied by the general community of biological scientists. At the same time, biological researchers are not inclined to adopt new technologies unless they can really “see” how they will help them work with their data more effectively or give them new means to address research questions. As one neuroscientist working on tool development stated, people in my field don’t “want me to tell them about any tool that they can’t go and use right now. ... if they can’t go and use it tomorrow then they’re not going to believe in it. ... [biologists] are funded for what they can do today, they are not funded for what they can do 10 years from now” [C1B1, 4/29/2003]. The post-prototype work does not require the biological expertise of the neuroscientists on the collaborative team, and there are limited incentives for computer science collaborators to remain engaged, since there is little science left in it for them either. In some cases, neuroscience teams have brought in engineers at this point in the process. But, they have tended not to have the user orientation or the “service to science” perspective needed to make the systems responsive to the scientists’ concerns and needs. There is an immediate need for SCSs in many large research projects. In the long run, smaller institutions will find it cost effective to include SCSs on the staff to increase the productivity of scientists who are spread too thinly.

It is also a significant problem when scientists must “ramp up” in order to build the information architecture to support the systems they desire. In one illustrative case, a post-doctoral biologist was asked to develop an ontology of a human disease as part of a collaborative informatics project. He began by learning how the underlying concepts fit together with the current major hypotheses, a process where his biological expertise was quite valuable. He stated that “I spent most of my time trying to figure out what I would put in an ontology, if I knew how to put one together.” Then, when he turned to learning about ontologies, he realized that he didn’t know where to go or how to begin. As the director of the project noted, the post-doc was struggling with what would become a much bigger challenge as the project progressed.

This has turned out to be a problem for the people who actually have to provide the content, because they’re really not sure. You know, ... it’s one thing if you tell them, well here’s five relations and here’s the terms and this is what you need to do. It’s quite another to say now you need to design pathways and do all this other stuff. [C1B1 4/4/2004]

The prototype and ontology examples are representative of a range of problems and activities involved at different stages of informatics projects. While the actual work is related to the primary objectives of the project, it no longer adheres to career expectations for either the

biologists or the computer scientists and is more removed from the scientific research that engages them.

At present, neuroscientists and computer scientists are invested in producing new tools, but their time is not best spent on the day-to-day work of sustaining informatics development. Some of the tasks involved may be easily covered by computer programming staff, but much of the work is about information organization, management, and communication and coordination among team members and other end user scientists. Activities may include locating rich data sources for extraction, enhancing metadata schemes, federating data from multiple locations, improving usability of interfaces, and specifying how the system can be improved to work for a larger audience or work in conjunction with other related systems. Many information problems involve not just one system, but a growing constellation of inter-related or complementary systems that need to be interoperable.

Computers, telecommunications, and particularly the Internet have also changed the nature of the work of systematists, ecologists, entomologists and almost every other type of biologist (Heidorn 2002a, 2002b). This change is particularly profound in publishing and information dissemination which has led to the creation of a set of new tasks that must be performed in the institutions where they work. This change is typified in natural history museums, botanical gardens, and natural history surveys in every state. All of these institutions now keep their collections in databases. Many of these institutions would like to make some of this information available to the public through web interfaces. The Z39.50 data federation that was first developed for libraries in the 1970s was adapted to biological collections (ZBIG), and this eventually evolved into the current Darwin core and DigIR standards used by GBIF. Some larger institutions build their own information management systems. For example, Missouri Botanical Garden built the highly successful TROPICOS system. Commercial as well as OpenSource solutions to museum management have also become available such as Emu and Specify, respectively. Throughout this evolution, scientists spent valuable research time dealing with the technical issues of coding and distributing their data. Libraries and dedicated information processing staffs developed at the larger institutions to support both research operations and now “standard” data publishing operations, such as the US NBII portal or the global GBIF portal.

### 2.3 Scope and limitations of current bioinformatics programs

In reviewing the current bioinformatics programs nationally, including the other programs being developed on our own campus, it is clear that the focus is not on service to science in the broad sense, or on the kinds of situations described above. But, these examples represent real concerns for the ongoing development and management of scientific research, since each year NSF and NIH invest heavily in these kinds of informatics projects with the aim of improving the practice of research for scientists at large. But, it appears that in many cases the scientists most involved in informatics development benefit while the typical scientist is often unaware of new developments, or at least unconvinced that there is anything for them in the informatics movement.

Based on the recent “Survey of Bioinformatics Programs in the United States,” by Hemminger, Losi and Bauers (2005) and a review of websites of existing programs, it is evident that almost all of the existing programs are heavily focused on computation and computer science. A few address information organization in terms of formal representation but do not consider problems such as metadata, thesauri construction, or indexing, for example. Several schools offer human-computer interaction types of courses, but only a couple offer courses that address user needs, scientific communication, or human information seeking behavior.

One program that has some parallels with our SCS model is the Certificate of Specialization in Bioinformatics offered through SILS at UNC-Chapel Hill. They offer a slate of more traditional

LIS courses, such as information retrieval, organization of information, and management of information agencies, to supplement biostatistics, biology requirements, and another group of LIS based electives. However, their program is an add-on to a masters in either library science or information science. Our program builds on the strengths of our existing masters option but is part of a separate campus-level bioinformatics masters degree program, requiring no additional certificate work.

The other program of note is the Master of Biomedical Informatics program at the Oregon Health Sciences University, an innovator in medical informatics education. While its focus is on healthcare and biomedical settings, and therefore the curriculum is more specialized than what we expect to design, the program is comparable in its main objectives: 1) To provide students with a theoretical and practical understanding of the role of information [in biomedical settings], 2) To provide students with a sound basis for implementing, developing, maintaining, and managing information resources and systems [in health care], 3) To provide students skills in the management of [biomedical] information, technology, and decision making.

#### 2.4 Scientific communication focus

Our graduate program would train a different kind of professional, focusing on developing expertise to fill the roles we see lacking in bioinformatics research teams and in the coordination of information systems development in the biological sciences. The overarching focus of the program will be on scholarly, or more specifically, scientific, communication. In the discourse of higher education, discussions of scholarly communication often cover only the new ways in which research papers are being published and disseminated. And, there is no doubt that electronic publishing on the Internet is quickly changing the format and means of distribution of scholarly works. However, there are many other dimensions of the scholarly communication cycle that are also changing in important ways in terms of how information is identified and mobilized for research purposes. In this sense, we conceive of scholarly communication as the entire set of information activities involved in scholarly exchange, including how information is found, integrated, and disseminated to produce new research results and scholarly products (Palmer, in press). Therefore, under the main rubric of scientific scholarly communication, this program will concentrate on training professionals to support science by building expertise in three areas:

- 1) Evaluation and implementation of information systems  
--user based assessment and continual quality improvement for the development of tools that work and are used.
- 2) Information acquisition, management, and dissemination.  
--development of digital libraries, data archives, institutional repositories, and related tools.
- 3) Information organization and integration  
--structuring information for optimal use and sharing, and standards development.

SCSs will develop functional applications that are integrated with current science practice. They will also attend to the more global concerns such as standards development, data and literature federation, and equitable dissemination of research results.

### **3. Building on the foundation of librarianship**

#### 3.1 Informationists

Our conception of the SCS is an extension of the informationist, a movement that began over 30 years ago as clinical medical librarianship (CML) and has now advanced beyond the clinical realm into scientific research teams. There has long been a disconnect between biomedical research, clinical practice, and healthcare provision. Despite recent developments in information technology and the pervasiveness of concepts like evidence-based medicine, which explicitly call

for the integration of research evidence with patient care, the knowledge that resides in medical and health-related journals, databases, and other resources often goes unused.

As Cimpl (1985) explains, librarians saw both an opportunity to meet clinical information needs and a place for themselves in the interdisciplinary rounding healthcare teams that emerged in the 1970s. CML services were offered “to provide information quickly to physicians and other members of the healthcare team; to influence the information-seeking behavior of clinicians and improve their library skills; and to establish the medical librarian’s role as a valid member of the healthcare team” (23). More recent CML programs still share many of these same objectives. For example, CML at Eskind Biomedical Library at Vanderbilt University Medical Center involves training and the development of deep clinical knowledge to both support interaction with rounding teams and to effectively search and interpret the biomedical literature (Giuse et al., 1998). Librarians in this program are promoted as “specialists who inhabit the intersection between an understanding of controlled vocabularies and a knowledge of medical concepts, and who provide a unique, value-added service which clinicians cannot necessarily duplicate on their own” (414). CML also overlaps with evidence-based medicine; both emphasize the emergence of questions, and the subsequent selecting, appraising, and applying of evidence to practice (Lipscomb, 2000).

Davidoff and Florance (2000) noted the numerous and wide-ranging barriers to the use of biomedical and health information and argued that new professionals were needed to link biomedical knowledge, literature, and clinical practice. They proposed informationists as a solution to the problems of literature scatter, the shortcomings of electronic indexing and searching, the limited training that healthcare professionals have in information retrieval, the time consuming nature of finding and selecting information, problems judging quality, extracting key pieces of information, and applying that information to particular clinical cases. Armed with knowledge of both information science and clinical work, informationists would function as members of clinical care teams and retrieve, synthesize, and present biomedical information.

More recently, there has been discussion over how the informationist concept, originally developed with clinical settings in mind, might apply to the research arena (Shipman et al., 2002). However, in certain realms of research, information professionals (albeit with different titles) have been supporting teams of scientists for some time (Neway, 1985). One of the earliest of such information services was located at the Battelle Memorial Institute, where information specialists worked together with subject experts in fields like organic chemistry, physics, engineering, and agricultural science to produce bibliographies “aimed toward solving a research problem as opposed to a random collection of references vaguely covering a technical subject” (Fizette et al., 1958, 253). Neway (1982) described how an information scientist with a background in the biological sciences became part of four research groups in a university microbiology department and helped the group members develop their personal journal article collections, exposed them to new journal titles, reduced their dependency on stated core journals, and freed up their time to pursue research. Even in more recent studies, researchers still describe a need for assistance in sifting through, sorting, and managing information. Scientists in a study by Grefsheim et al. (1991) wanted someone to train them to more effectively identify, use, and organize information. They identified librarians as the “agents” who could bring order to available biomedical information resources. Similarly, the University of Washington Health Sciences Libraries developed a heavily utilized bioinformatics services program complete with consultation services, education, training, and other resources to enhance access for researchers to various biological information resources (Yarfitz & Ketchell, 2000).

At present, there are more information resources available to biomedical researchers than ever before, and countless more are in development. They range from the bibliographic to the nonbibliographic, and include Internet websites, data analysis software, visualization tools, and

web-based databases containing published literature, DNA and protein sequences, and other kinds of content. The professionals who work in collaboration with teams of scientists to facilitate their interaction with and use of these resources may come from either information or health-related backgrounds (Detlefsen, 2002). But regardless, as Florance et al. (2002) explain, preparing information specialists to work in such “information-rich environments and to participate as peers in problem solving” requires cross training in library and information science and discipline knowledge of scientific domains, and their training should include an internship in a practice setting. We agree that the expertise of SCSs working in the contemporary biological research environment will need to span the scientific research domain and information science.

### 3.2 Core knowledge

For admission to the program students will need to have a solid science background at least at the undergraduate level or be willing to remediate with coursework in our biological departments. In the information specialists program they will be trained to work on scientific teams or in support roles in organizations such as academic departments, centers or institutes, companies, research libraries, and museums. As an LIS school, our approach to curriculum development and course content will represent core LIS principles including service, equitable access and dissemination, and diversity of resources. It will also reflect our long history in research librarianship where professionals have been responsible for large-scale monitoring, coordination, and access to scholarly and scientific intellectual property, expert information searching, and consortial resource development. LIS is the only field that is concerned with the full landscape of scientific information and the interactions therein, and with provision of services to exploit that base of knowledge (Bates, 1999; White, Bates, and Wilson, 1992).

As outlined in section 2.4, the program will concentrate on training professionals to support science by building expertise in three areas that we conceive of as falling under the broader rubric of scientific communication:

- Evaluation and implementation of information systems
- Information acquisition, management, and dissemination
- Information organization and integration

Each of these thrusts is important to a number of application areas, including informatics tool development, collaboratory design, digital library and museum collections, open access digital publishing, data curation, long-term storage and management, and standards development and implementation. These areas of expertise require core knowledge that is partly covered by some GSLIS masters courses. Where possible, courses will be revised to accommodate the needs of SCS students. For example, we can build on these existing courses:

Representing and Organizing Information	Interfaces to Information Systems
Building Digital Libraries	Indexing and Abstracting
Health Sciences, Information Services and Resources	Architecture of Networked Information Systems
Information Sources and Services in the Sciences	Implementation of Information Retrieval Systems
Use and Users of Information	Electronic Publishing
Document Modeling	

We expect to develop completely new courses in the curriculum to encompass:

Scientific data and procedure standards	Sociotechnical perspectives on scientific practice
Scientific classification and vocabulary	State-of-the-art informatics resources and tools

Ontology development	Scientific literatures and bibliometrics
Data curation and long-term data management	Open access repositories
Discovery informatics and data mining	Project management
Interdisciplinary scientific collaboration	

#### **4. Infrastructure and resources**

##### 4.1 Academic base

While information technologies are quickly changing, many of the roles we have identified for SCSs are natural extensions of the kind of training traditionally provided by LIS programs. We have always trained librarians to be science subject specialists in research libraries or special librarians in companies that perform scientific research, such as in the pharmaceutical industry. These librarians have responsibilities for managing and promoting effective information use for the benefit of research and often competitive advantage. Now we need to prepare more information specialists to work on research teams and in research institutions. The new dimensions and complexities of contemporary research need to be more fully integrated into our curriculum, and, we would argue, into curricula in peer schools across the country.

Because of the new campus bioinformatics program, this is the optimal time for GSLIS to concentrate on curricular development in this area. The campus program will manage a set of core courses in biology and computer science from which our students will draw for 2 of the 9 courses required by the campus criteria. Moreover, the new certificate of advanced study in digital libraries to begin in the fall of 2005 has laid the groundwork for developing more technologically based courses. Many of the courses developed for the digital libraries students can be readily customized for biological informatics applications. In biological informatics, we have a health informatics course in place and a biodiversity informatics course in development. Another course on information transfer and collaboration in science is also slated for development in the coming year. These will be the first electives available to the students. The others will be developed during the three year period covered by this proposal.

##### 4.2 Key personnel

The PIs have significant research, teaching and outreach experience in biological informatics systems. Dr. Palmer has worked closely with scientists to learn about the information systems and services needed to promote discovery and interdisciplinary inquiry (Palmer 1996, 1999, 2001; Palmer, Cragin, and Hogan 2004). This research applies directly to current issues in scholarly communication (2004, in press), and relates to her teaching foci in digital collection development and use and users of information.

Dr. Heidorn has run a series of projects in electronic publishing of flora and fauna (Heidorn, 2001, 2003b; Cui & Heidorn, 2002), computer-based biological data gathering in the field (Heidorn et al., 2002a, 2002b), interactive identification key development, and natural history museum specimen access. All of these projects include substantial educational experiences for both undergraduates and graduate students in many departments. Some have included a substantial life-long learning component as well, where adult non-university participants learned about both biology and computer technology.

Other faculty members at GSLIS who will be involved in the curriculum development are Dr. Bruce Schatz and Dr. Linda Smith. Bruce Schatz is working closely with the campus bioinformatics committee, is core teaching faculty in GSLIS in the information systems and health informatics areas, and has managed many successful large-scale information systems development projects in the sciences (i.e., Schatz, 1991; Schatz, Mischo, Cole, et al., 1999). Linda C. Smith is Professor and Associate Dean in GSLIS, and regularly teaches the course in Information Sources and Services in the Sciences. As a member and past chair of the American

Association for the Advancement of Science Section T (Information, Computing and Communication), she has a longstanding interest in preparation of students for emerging roles as scientific communication specialists (Smith, 1987).

#### 4.3 Engagement with the scientific community

All of the proposed curriculum development must be done in close coordination with practicing scientific communities and with a solid understanding of the aims of biological research that information systems are designed to support. Unlike many information technology jobs, SCS work will require significant knowledge of the biological domains served. In our program, students will not only gain a broad understanding of scientific communication and information organization, retrieval, and management, but they will also gain a strong understanding of how informatics fits within the biological science disciplines. Therefore, it is essential that practicing research scientists guide how these professionals will be trained. We will involve scientists from several disciplines of biology and many institutions in the SCS program. They will participate as part of the scientific advisory board, as internship supervisors, and as adjunct or guest lecturers. Through our local and national collaborations we also expect to identify new and useful research projects which would be difficult without the collaboration on teaching and internships, and continue our ongoing work toward **Objective 5**—expand understanding of the role of informatics in scientific progress.

##### 4.3.1 Local Expertise

There are many scientists and research units at the UIUC that will participate in the SCS program. UIUC is one of the leading research universities in the country and has a strong focus on biological sciences. In addition, the University has the distinction of being co-located with all of the State Surveys. In the context of the SCS program, the scientists in the Illinois Natural History Survey are the most relevant. Professors, scientists, and other professionals in these units will be able to have regular contact with our students and GSLIS research and teaching faculty.

Stephen Downie, Sydney Cameron, and James Whitfield in UIUC Integrative Biology, Plant Biology, and Entomology teach courses in the Principles of Systematics that are part of the GSLIS requirements for the bioinformatics degree. Ken Robertson at the Illinois Natural History Survey (INHS) and with appointments in the UIUC Department of Natural Resources and Environmental Systems will help teach and supervise students. Chris Dietrich is an entomologist and curator at the INHS Museum and has a joint appointment at UIUC. He is actively involved in information systems work and can serve as an advisor to GSLIS on both biology and IT issues.

All of these individuals and others will be able to direct students to the SCS program, and some of the faculty listed above have students who would like to begin such a program now. Some students may register for the masters degree in bioinformatics. On the other hand, many institutions can not afford to hire full time technical and library staff but will continue to rely on scientists to perform this SCS tasks. Students pursuing training in masters and PhD programs in biology will be able to take courses in our program to learn skills that they will need on the job.

##### 4.3.2 Cooperating National Institutions

Many scientists and other professionals have already developed many of the skills that we need to teach to the students in our masters programs. They work in some of the nation's leading biological research institutions, including the Smithsonian Institution, the American Museum of Natural History, Missouri Botanical Garden, the Peabody Museum at Yale, Biomedical Informatics Research Network, The Organization for Tropical Studies and Biomedical Informatics Research Network. All of these institutions run large scale biological informatics programs that would serve as ideal areas of study and examples of best practices for our students. We expect to work closely with these and other institutions. The operations of a few are outlined below.

The Missouri Botanical Garden (MBG) is one of the world's top botanical research and conservation institutions. The Garden's dozens of Ph.D. researchers work to strengthen scientific expertise in developing countries to protect and manage biodiversity before it's too late. MBG is a leader in information technology with many cutting edge projects including for example, TROPICOS, the world's largest database of plant information, contains fully web-searchable records for over 900,000 plant names and nearly 2 million specimens. Over 50,000 plant images are also linked to their records in TROPICOS.

The Peabody Museum Informatics Program at Yale collaborates on two research projects with Bryan Heidorn. This includes the NSF funded Herbis project (<http://www.herbis.org/>) and the Moore BioGeomancer Project. Reed Beaman is a plant biologist and runs the Peabody Informatics program. He will be able to serve as an excellent advisor for the SCS program and the center could serve as a potential site for internships for our students.

The Biomedical Informatics Research Network (BIRN) consortium currently involves 26 research sites from 19 universities and hospitals that participate in one or more of three test bed projects: Morphometry BIRN, Function BIRN, and Mouse BIRN. These projects are centered on structural and/or functional brain imaging of human neurological disorders and associated animal models of disorders including Alzheimer's disease, depression, schizophrenia, multiple sclerosis, attention deficit disorder, brain cancer, and Parkinson's disease. Our collaborator is Maryann Martone at the BIRN Coordinating Center at the University of California at San Diego ([www.nbirn.net/TestBeds/CoordinatingCenter/index.htm](http://www.nbirn.net/TestBeds/CoordinatingCenter/index.htm)).

## **5. Plan of work and project management**

The four primary objectives will be met through a range of activities, many of which will be ongoing processes throughout the course of the project.

**Objective 1.** Develop curriculum for the specialization that builds logically on existing graduate programs at GSLIS and the core provided by the new campus-level bioinformatics degree.

### **Activities:**

- Establish an advisory board for biological informatics education. Our advisors will be scientists and other professionals on the front lines of biological informatics development. We will begin with the partners identified above and add members to round out representation across appropriate scientific fields. We will hold a two-day summit at the beginning of Year One to establish guidelines and goals for program development over the course of the project. After that, semi-annual meetings will be held via conferencing technology. The Board will reconvene at the end of the project to assess overall progress, outcomes, and advise on continued development of the program.
- Conduct a needs assessment of biological informatics expertise, targeting research departments, labs, and biological informatics initiatives across the country. This work will include a survey of research scientists and project managers and will be followed up by phone interviews with respondents who volunteer to provide more in-depth information on how new SCS professionals can contribute to research operations.
- Monitor the current job market, by continuing to build our collection of position announcements and job descriptions from major science journals and websites. Send a representative to conferences, such as American Medical Informatics Association, International Conference on Bioinformatics and its Applications, and Preservation of Natural History Collections.

- Update our preliminary review of curriculum materials from bioinformatics, chemical informatics, and genomics programs. Consult with librarians who are connected with current bioinformatics programs.
- Refine and develop curriculum. Review existing courses and coordinate with the current DL-CAS program committee to evolve courses that meet the needs of both programs. Design a foundations course and three new courses in the core areas, based on the outcomes of activities listed above. The foundations course will provide an introduction to the new core knowledge areas identified in section 3.2. The three area courses will build on that foundation to provide depth in the three areas identified in section 2.4, cutting across the biological domains to prepare students for the range of local and global scientific information work.
- Evaluate courses and program through the UIUC course evaluation system and a survey of graduating students on the effectiveness of the program. We also plan to track placement and do follow-up evaluation with employers one year after placement.

**Objective 2.** Establish internships at institutions and laboratories where students can develop and apply their growing expertise.

**Activities:**

- Build partnerships with institutions that will offer internship opportunities for students to work with bioinformatics leaders. Collaborate with site supervisors to establish internship guidelines and outcome measures. Continue to explore and build relationships with other internship sites.
- Develop an evaluation process to assess the performance of interns and to assess the value of individual internship positions.

**Objective 3.** Develop mechanisms for integrating course work into current informatics research at GSLIS, other UIUC departments, and partner institutions.

**Activities:**

- Develop case-based modules of study and assignments in the three core areas in coordination with GSLIS and UIUC research projects and those at partner institutions.
- Identify opportunities for student projects to contribute to ongoing research projects.

**Objective 4.** Share our educational approach to biological informatics with other schools interested in developing similar specializations.

**Activities:**

- Develop best practices reports on the education of biological informatics professionals. Make all project documents, syllabi, lecture materials, and reports available on the project web site.
- Disseminate progress and outcomes of the project at information science, biology, and science education conferences.

	Year 1	Year 2	Year 3
<b>Objective 1</b>			
Advisory Board	Install board Program Summit (February) 6-month meeting (June)	Semi-annual meetings (Jan. & June)	Semi-annual meetings (Jan. and June) Program Review Meeting (Dec.)

Build Curriculum	Review Curriculum Needs Assessment Monitor Job market		Evaluate courses and program
<b>Objective 2</b>			
Establish Internships	Build institutional partnerships Establish site guidelines and outcome measures	Develop Intern review and evaluation process (intern/site)	
<b>Objective 3</b>			
Integrate coursework into research		Coordinate/ develop case-based modules Identify student project opportunities	
<b>Objective 4</b>			
Share Educational Approach	Develop program web site Develop best practices reports	Maintain website; add materials Disseminate progress and Outcomes	

Targets: The new program is slated to begin in the fall of 2006, but elective courses will not be needed until spring of 2007. We expect to allow up to 30 new students per year and that time to graduation will be approximately 3-4 semesters, depending on the student's scientific background.

The co-PIs will co-direct the program development, providing ongoing oversight and guidance, and they will be responsible for developing advisory and internship partnerships. In the development of course content, the PIs will work closely with other GSLIS and partner faculty to assure coverage of core knowledge, to integrate the expertise of working scientists and informaticists, and to integrate current research into class experiences. A doctoral research assistant will be recruited to coordinate the project and perform other activities outlined above.

Preliminary advisory board:

Thomas Garnett, Assistant Director for Digital Library and Information Systems, Smithsonian Institution Libraries

John Kress, Chairman and Curator Department of Botany, Smithsonian Institute

Maryann Martone, Scientific Coordinator, Biomedical Informatics Research Network, University of California at San Diego

Chuck Miller, Chief Information Officer, Missouri Botanical Garden

Tom Moritz, Boeschstein Director, Library Services, American Museum of Natural History

Neil Smalheiser, Arrowsmith project director, Psychiatric Institute, University of Illinois at Chicago

## 6. Results from prior NSF work

Carole Palmer:

Information Work and Discovery Potentials in Neuroscience Research. NSF IIS-0222848. \$346,870, August 2002- July 2005. The project (IDN: Information and Discovery in Neuroscience) examines high impact information in the discovery process in neuroscience, with a focus on the potential of literature mining for supporting the integration of information from

different brain research communities. Field studies of 12 neuroscience research projects are being conducted, and we are working closely with the developers of Arrowsmith (a NIH Human Brain Project grant), providing feedback to inform development of the tool they are developing for literature mining in PubMed. Data analysis is still ongoing on this project, but results with particular relevance for this proposal include: 1) researchers had pronounced difficulty locating specifics on protocol, instrumentation, measurement, and results; 2) retrospective, non-digital literature was often ignored; 3) background work outside one's specialization and assessing findings and project risks require the most time and effort; 4) routine, high reliance on PubMed suggests that literature mining could be widely used if well integrated with current functionality. However, it also raises questions about how much relevant information is being missed that is not covered in PubMed (i.e., nutritional, environmental) and how to best integrate complementary databases. 5) Researchers are not able to manipulate scientific information at a suitable level of granularity, which relates to other problems with standardization and curation of data and the labor involved in data extraction for building knowledge bases and evaluating their effectiveness. The project has supported two doctoral research assistants, and one is currently conducting her thesis research on scientific data curation, archiving, and sharing, as a direct extension of initial findings on this project. Publications and presentations include:

- Palmer, C. L. (Forthcoming). Scholarly work and the shaping of digital access. *Journal of the American Society for Information Science and Technology*.
- 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 and Technology annual meeting, Providence, RI, November 13-18, 2004*, pp. 448-455.
- Palmer, C. L. (2004). Boundary work and doable science. *Association for Computing Machinery, Grace Hopper Celebration, Chicago, Illinois, October, 2004*.
- Palmer, C. L., Cragin, M. H. and Hogan, T. P. (2003). Information and discovery in neuroscience. *American Society for Information Science and Technology (ASIS&T) annual meeting, Long Beach, CA, 19-22 October, 2003*, pp. 540-541.

Bryan Heidorn:

P. Bryan Heidorn (PI) and Michael Jeffords (CO-PI). \$287,042, January 1, 2000 - May 31, 2002. NSF/BIO/Database Activities: Biological Information Browsing Environment: Visual browsing environment with faceted retrieval of full text, similarity matching and knowledge extraction. [<http://www.biobrowser.org>] This is a project in automatic flora document structuring. While many species go unnamed and undescribed, natural scientists have spent over two hundred years cataloging and describing species. This rich body of texts resides in libraries throughout the world. A challenge for the creation of digital libraries of documents like these is to enhance the value of paper-based publications by providing digital access to the materials. The process includes automatic text segmentation, automated XML markup, structure-based indexing, thesaurus extraction for query expansion and on-line definitions. The Biological Information Browsing Environment team (<http://www.biobrowser.org>) includes over ten paid graduate students and numerous student projects. The students developed text processing, machine learning and pattern matching tools, the structures and indexes to the Flora of North America (FNA). We introduce taxonomically accurate hyperlinking between species, genera and families of plants. We process the Categorical Glossary of FNA and the Species Plantarum to provide inline definitions and automatic query expansion for synonyms, broader and narrower terms. We developed an XML aware full-text indexing engine to allow users to search parts of the documents such as the nomenclature, the morphological description and natural history independently (Heidorn, 2000). This project led to a project funded by IMLS for creating XML formatted species descriptions and searching tools. The NSF grant led to the dissertation of Hong Cui, now a professor of information science at the University of Western Ontario and the data sets are the basis for the soon to be defended dissertation of Xiaoya Tang, now a faculty member at Emporia State University. Lei Ding produced a masters thesis on parsing and structuring a

floristic glossary. Seven graduate students worked on different aspects of this project either as RAs, independent studies or other thesis or dissertation work. The biological information text extraction and formatting work led into the newly begun Betty and Gordon Moore grant: Biogeomancer to develop text processing tools to help georeference the billions of natural history specimen in museums world wide (an excellent working application for our students and graduates). It also led to the formatting of museum label data in the ongoing NSF BDI-0345387 titled "Collaborative Research: Rapid Digital Specimen Image and Data Capture: A Web Services Solution ". Publications and presentations include:

Heidorn, P. Bryan. (2003) OpenKey: Illinois-North Carolina Collaborative Environment for Botanical Resources. , *First Monday*, 8(5) (May 2003). [<http://firstmonday.org/>]

Heidorn, P. Bryan, Bharat Mehra, Mary Lokhaiser. (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*, 53, (14), 1251-1258.

[<http://www.isrl.uiuc.edu/~pheidorn/papers/Submission3.pdf>]

Heidorn, P. Bryan. (2001) A Tool for Multipurpose Use of Online Flora and Fauna: The Biological Information Browsing Environment (BIBE), *First Monday*, 6(2) (February 2001). [<http://firstmonday.org/>]

Heidorn, P. Bryan (2002). Biodiversity and Biocomplexity Informatics: Policy and Implementation Science versus Citizen Science. 2002 Joint Conference for Digital Libraries, July 14-17, Portland OR. p. 362-364.

Hong, Cui & Heidorn, P. Bryan (2002). An Approach to Automatic Classification of Text for Information Retrieval. 2002 Joint Conference for Digital Libraries, July 14-17, Portland OR.

Heidorn, P. Bryan. Beyond Paper on the Web: Highly Functional Flora. International Symposium on Plant Diversity in East Asia. March 5-6, 2003. Taichang, Taiwan.

## References

- Bates, M.J. (1999). The invisible substrate of information science. *Journal of the American Society for Information Science*, 50(12), 1043-1050.
- Cimpl, K. (1985). Clinical medical librarianship: A review of the literature. *Bulletin of the Medical Library Association*, 73(1), 21-28.
- Convention on Biological Diversity (1992). *Convention on biological diversity*. Retrieved April 25, 2005, from <http://www.biodiv.org/doc/legal/cbd-en.pdf>
- Cui, H. & Heidorn, P. B. (2002). An approach to automatic classification of text for information retrieval. 2002 Joint Conference for Digital Libraries, July 14-17, Portland OR.
- Davidoff, F. & Florance, V. (2000). The informationist: A new health profession? *Annals of Internal Medicine*, 132(12), 996-998.
- Detlefsen, E.G. (2002). The education of informationists, from the perspective of a library and information sciences educator. *Journal of the Medical Library Association*, 90(1), 59-67.
- Fizette, M.Y., Jones, B.E. & Gibson, R.W. (1958). The bibliographic research team. *Special Libraries*, 49, 253-255.
- Florance, V., Giuse, N.B., Ketchell, D.S. (2002). Information in context: Integrating information specialists into practice settings. *Journal of the Medical Library Association*, 90(1), 49-58.
- Giuse, N.B., Kafantaris, S.R., Miller, M.D., Wilder, K.S., Martin, S.L., Sathe, N.A. & Campbell, J.D. (1998). Clinical medical librarianship: The Vanderbilt experience. *Bulletin of the Medical Library Association*, 86(3), 412-416.
- Grefsheim, S., Franklin, J. & Cunningham, D. (1991). Biotechnology awareness study, part 1: Where scientists get their information. *Bulletin of the Medical Library Association*, 79(1), 36-44.
- Heidorn, P. B. (2000). The interaction of result set display dimensionality and cognitive factors in information retrieval systems. Annual Conference of the American Society for Information Science, November 1-5, Chicago, IL.
- Heidorn, P. B. (2001) A tool for multipurpose use of online flora and fauna: The Biological Information Browsing Environment (BIBE), *First Monday*, 6(2).  
[<http://firstmonday.org/>]
- Heidorn, P. B, Mehra, B., Lokhaiser, M.. (2002a). 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*, 53(14), 1251-1258.

[<http://www.isrl.uiuc.edu/~pheidorn/papers/Submission3.pdf>]

Heidorn, P. B. (2002b). Biodiversity and biocomplexity informatics: Policy and implementation science versus citizen science. 2002 Joint Conference for Digital Libraries, July 14-17, Portland OR, pp. 362-364.

Heidorn, P.B. (2003a). Biological informatics: A comparison of biodiversity informatics and neuroinformatics. *Bulletin of the American Society for Information Science and Technology*, 30(1), 12-13.

Heidorn, P. B. (2003b) OpenKey: Illinois-North Carolina collaborative environment for botanical resources, *First Monday*, 8(5). [<http://firstmonday.org/>]

Hemminger, B.M., Losi, T. & Bauers, A. (2005). Survey of bioinformatics programs in the United States. *Journal of the American Society for Information Science and Technology*, 56(5), 529-537.

Kling, R., and McKim, G. (2000). Not just a matter of time: Field differences and the shaping of electronic media in supporting scientific communication. *Journal of the American Society for Information Science*, 51(14), 1306-1320.

Lipscomb, C.E. (2000). Clinical librarianship. *Bulletin of the Medical Library Association*, 88(4), 393-395.

Neway, J.M. (1982). The role of the information specialist in academic research. *Online Review*, 6, 527-535.

Neway, J.M. (1985). *Information specialist as team player in the research process*. Westport: Greenwood Press.

Palmer, C.L. (1996). Information work at the boundaries of science: Linking information services to research practices. *Library Trends*, 45(2), 165-191.

Palmer, C.L. (1999). Structures and strategies of interdisciplinary science. *Journal of the American Society for Information Science*, 50(3), 242-253.

Palmer, C.L. (2001). *Work at the boundaries of science: Information and the interdisciplinary research process*. Dordrecht: Kluwer Academic Publishers.

Palmer, C.L. (in press). Scholarly work and the shaping of digital access. *Journal of the American Society for Information Science and Technology*.

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

Schatz, B. R. (1991). Building an electronic community system. *Journal of Management Information Systems*, 8(3), 87-101.

Schatz, B.R., Mischo, W. H., Cole, T. W. et. al. (1999). Federated search of scientific literature. *IEEE Computer*, 32(2), 51-59.

Shipman, J.P., Cunningham, D.J., Holst, R. & Watson, L.A. (2002). The informationist conference: Report. *Journal of the Medical Library Association*, 90(4), 458-464.

Smith, L.C. (1987). Education for sci-tech librarianship: Retrospect and prospect. *Science & Technology Libraries* 8(1), 75-88.

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

Wilson, E.O. (2000). A global biodiversity map. *Science*, 289(5488), 2279.

Yarfitz, S. & Ketchell, D.S. (2000). A library-based bioinformatics services program. *Bulletin of the Medical Library Association*, 88(1), 36-48.