



# Supporting Biological Information Work: Research and Education for Digital Resources and Long-lived Data

Carole L. Palmer, Melissa H. Cragin, P. Bryan Heidorn, Dan Wright

Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign  
501 E. Daniel St., Champaign, IL 61820

## 1. Introduction

New practices are emerging in all stages of biological research, from data collection through dissemination of results. Through a series of cooperative projects with biologists working in data-intensive and informatics-based domains, we have documented requirements for digital libraries, tool development, and data management techniques to support contemporary scientific practice. This research is now serving as the foundation for a new biological informatics master's program to train scientific information specialists to manage and integrate scientific information and tools to support scientific problem solving and communication.

## 2. Biological Information Expertise

Areas of information support that have emerged as priorities to complement the expertise of biological and computer scientists include digital library and repository development, data curation and preservation, ontology and standards development for interoperable systems, and literature-based discovery. To respond to the qualitative changes in biological research and the specific workforce gaps identified in our research, we are developing a comprehensive master's level training program in scientific communication, as part of a campus-wide bioinformatics initiative at the University of Illinois at Urbana-Champaign. The program will train a new generation of Library and Information Science professionals to serve in scientific research environments.

## 3. Results from prior work that inform the development of the SCI program:



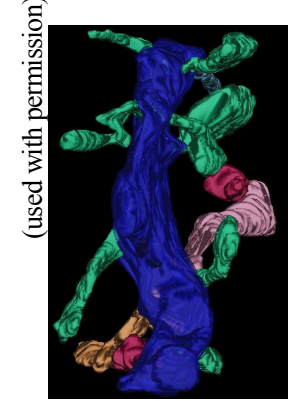
### Biodiversity Research and Development Projects

- ▶ Many groups that collect biodiversity data in the field do not have the skill sets or the technical infrastructure to make the data available to others. Scientific Communication Specialists can help to solve this problem.
- ▶ The museums of the world contain between 2 and 3 billion specimens, but Internet-accessible electronic catalogs exist for only 5% of these collections. Informaticians are helping to solve this problem by developing international data interchange standards, collections digitization tools, and other data acquisition and distribution frameworks.
- ▶ The Natural history literature stretches back hundreds of years and is critical to biologists work in systematics, ecology and other fields. Digitizing and connecting these materials to on-line resources such as PubMed and Genbank are a major challenge for the informatics and biology community.



### Information and Discovery in Neuroscience

- ▶ There is a tendency for biology-computer science collaborations to languish once a prototype informatics system is produced; work is required to get the system past the prototype.
- ▶ There is significant "ramp up" required for scientists who need to build information architecture to support their research. While their domain expertise is requisite, there is a great challenge for scientists who don't know where or how to begin to develop knowledge systems such as lexicons or ontologies.
- ▶ We have identified a continuum of weak and strong information work activities that has implications for the prioritization of new information systems and services to support scientific research.

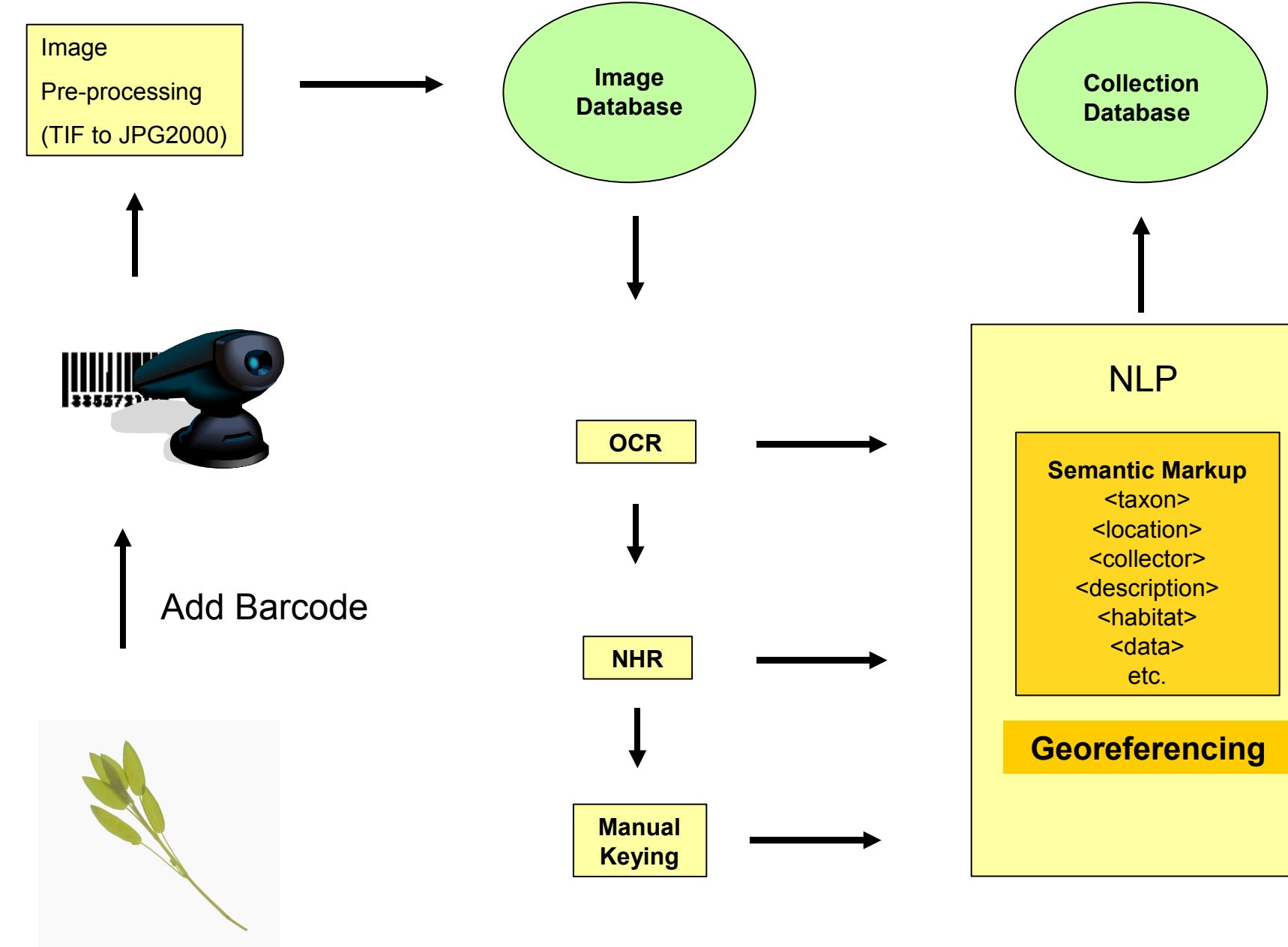


### Shared Scientific Data Collections and Scholarly Communication

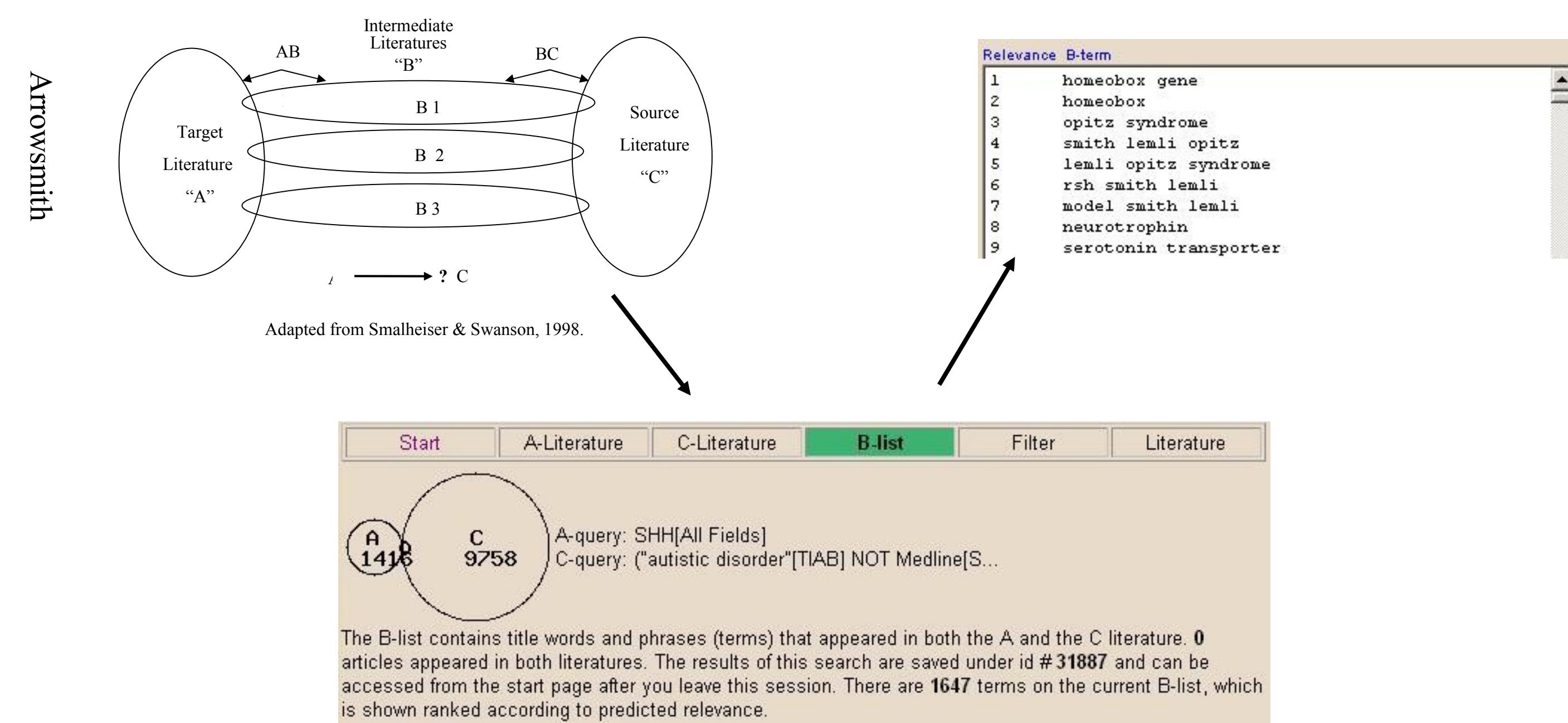
- ▶ The lack of standards for citing biological data or data sets poses problems for both the depositing scientists and anyone seeking to re-use the data. Depositing scientists, or authors, are concerned that they will not be credited for their research products. End users need know *how* to cite data and collections they use.
- ▶ Dynamic and federated searches will return data drawn together from different collections or databases – what is the best way to design databases to include sufficient administrative metadata?
- ▶ How will data curation infrastructures intersect scholarly communication processes?

## Examples of Biological Information Problems Addressed in Our Research

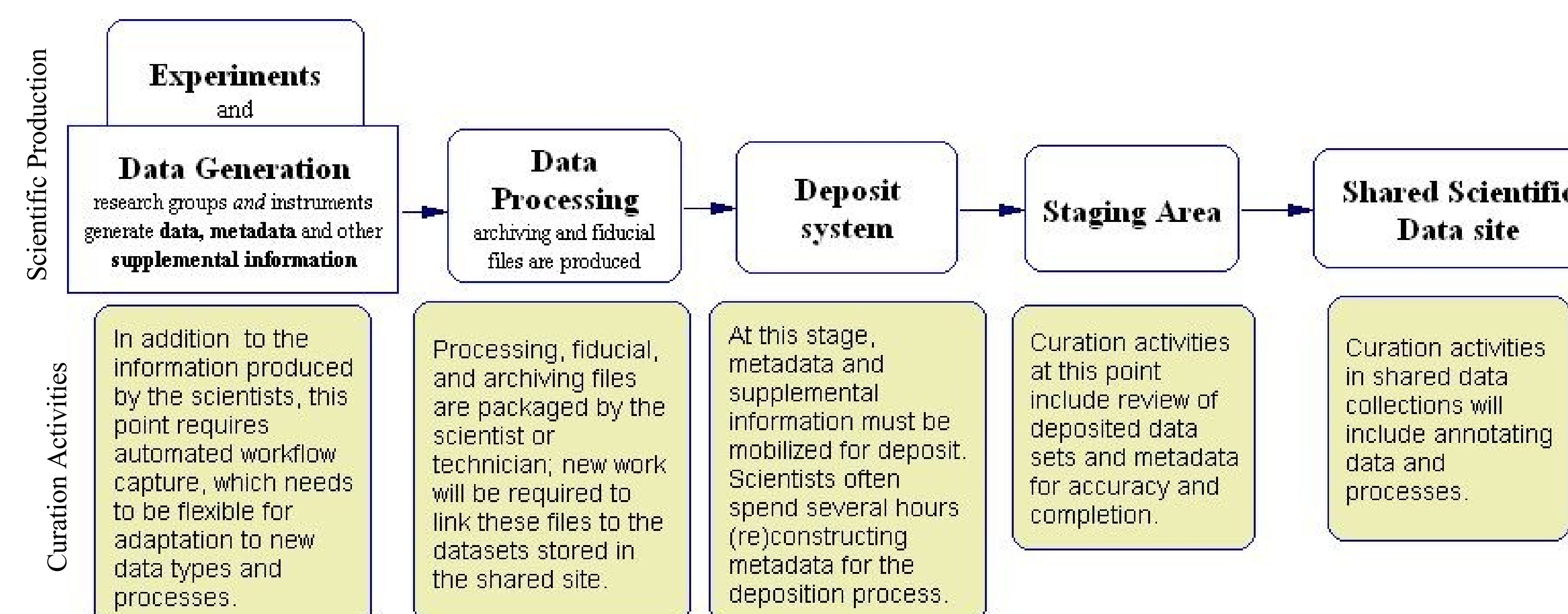
1. HERBIS Work Flow: This "proof of concept" project is an initial implementation of 'one-button' specimen imaging and data capture.



2. Literature based discovery (LBD) techniques are being developed by the Arrowsmith Project to promote discovery through linking findings from disconnected literatures.



3. Data curation activities involved in depositing biological data for shared use are problematic for scientists. These activities are new and there are no standards for integrating them into the daily and routine work of their research.



This project has been supported by Information and Discovery in Neuroscience, NSF-0222848 (CL Palmer, PI), and A Graduate Program for Scientific Communication Specialists: Getting Past the Prototype in Biological Informatics, NSF-IIS-0534567 (C.L. Palmer and P.B. Heidorn co-PIs).

## 4. Training Program for Biological Information Specialists (BIS)

Specifically, the program provides applied skills in building and evaluating systems that mediate effectively between users and collections, and emphasizes the range of library and information science including: collection development, classification schemes, information retrieval, knowledge representation, user evaluation, data curation, and policy standards. Our students are taught to develop information management systems in biological applications, with opportunities to consider a broad spectrum of domains including molecular biology, environmental ecology, and biomedicine.

The degree requires a total of 36 hours of coursework. Students will arrange custom programs of study, suitable for the information management of their particular bioinformatics application. For more information on the MS Program, including course requirements and electives, please see: [www.lis.uiuc.edu/programs/ms-bioinformatics.html](http://www.lis.uiuc.edu/programs/ms-bioinformatics.html)

The program provides students with access to international experts from across the University who specialize in many areas of biology and information management including information science, bioinformatics, biology, chemistry, statistics, and computer science.

## 5. Cooperating 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, and they work in some of the nation's leading biological research institutions. These same institutions run large scale biological informatics programs that will serve as ideal areas of study and examples of best practices for our students. We are partnering with those listed below, as well as other institutions. The operations of a our primary collaborative sites are outlined below.

The **Missouri Botanical Garden** (MBG) is one of the world's top botanical research and conservation institutions. 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. ([www.mobot.org/default.asp](http://www.mobot.org/default.asp))

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. ([www.nbirn.net/TestBeds/CoordinatingCenter/index.htm](http://www.nbirn.net/TestBeds/CoordinatingCenter/index.htm))

The **Smithsonian Institution** in Washington, D.C., annually hosts 5.4 million visitors and over 11 million visitors to the National Museum of Natural History website. The year 2005 saw the publication of over 500 research articles, including 19 in high impact journals, such as Science and Nature. Their Collections and Research Information Systems (CRIS) is a distributed, multimedia system supporting the documentation, management, analysis, and delivery of collections, educational, and research resources held and produced by the Museum. ([www.mnh.si.edu/](http://www.mnh.si.edu/))

## 6. 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
- **Christie Stephenson**, Acting Director, Darwin Digital Library of Evolution, American Museum of Natural History
- **Neil Smalheiser**, Arrowsmith project director, Psychiatric Institute, University of Illinois at Chicago

## 7. References

- [1] See for example, National Science Board (September 2005): NSB-05-40, Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century. <http://www.nsf.gov/pubs/2005/nsb0540/>
- [2] Projects include Carole Palmer's Information and Discovery in Neuroscience (NSF IIS-0222848), Bryan Heidorn's Georeferencing Museum Specimen Sources (Moore 2005-2929-00), and Heidorn and Palmer's, Internet Environment for BioDiversity Survey Collaboration and Verification (NSF BDI-011391).
- [3] Smalheiser, N. R. & Swanson, D.R. Using ARROWSMITH: A computer-assisted approach to formulating and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine*, 57(3), 149-153.