

The Role of Information Science in Gathering Biodiversity and Neuroscience Data

by Geoffrey A. Levin and Melissa H. Cragin

Geoffrey A. Levin is director of the Center for Biodiversity at the Illinois Natural History Survey, 607 E. Peabody Dr., Champaign, IL 61820-6970.

He can be reached at 217-244-7481 or by e-mail: glevin@inhs.uiuc.edu

Melissa H. Cragin is a doctoral student in the Graduate School of Library and Information Science at the University of Illinois, Urbana-Champaign, 501 E. Daniel St., Champaign, IL 61820.

She can be reached at 217-244-0866, or by e-mail: cragin@uiuc.edu.

As with the rest of science, biodiversity and neuroscience are becoming increasingly technological and data-rich. At the same time biodiversity studies retain many traditional tools and materials, while in neuroscience new tools are frequently developed. Information science can help both fields develop tools that exploit modern technologies to increase data-gathering efficiency, to improve quality control and, where necessary, to integrate historical and modern methods and materials. We will present overviews of biodiversity and neuroscience separately and conclude by summarizing how information science can help both fields address similar issues.

Biodiversity

A valuable perspective on biodiversity studies can be gained from considering the Lewis and Clark Expedition, whose bicentennial the United States will celebrate in 2004. Like explorers before and after them, they and their Corps of Discovery were charged not only with exploring and mapping the land they crossed, but also with documenting the plants, animals, minerals and indigenous peoples they encountered (Ambrose, 1996). They collected numerous specimens and artifacts that they removed from the region of origin and sent to institutions and individuals in the “developed” world, in this case the northeastern United States. Most of the plant specimens, for example, are at the Academy of Natural Sciences in Philadelphia (www.acnatsci.org/museum/lewisclark/l&c_herbarium.html). There, experts studied the specimens and described new species, often over a span of many years. Unlike most other 19th century explorers Lewis and Clark kept extensive notes that still survive, so it is pos-

sible to know fairly precisely where the specimens came from, their ecological setting and other relevant information. More typical are notes like those of mid-19th century explorer Charles Wright, many of whose specimens, distributed to museums in the eastern United States and Europe, are accompanied only by notes giving the month and year he collected and the location as “western Texas.”

Collecting practices like these continued well into the 20th century, persisting longest in tropical areas. In well-explored areas, collectors were able to give precise locations, but in tropical areas they may only have known the river they were along and how many hours of paddling it took to get there from a distant village. Ecological information with the specimens tended to be idiosyncratic, with some scientists providing good descriptions of vegetation, soil and other relevant conditions, whereas others provided little or no data. Beginning in the 19th century, the most significant change was the elimination of the amateur collector. Instead, predominant practice throughout the 20th century has been for professional scientists to serve both as field collectors and museum experts. Museums worldwide now hold an estimated two to three billion biodiversity specimens, about 75% in the industrialized countries, and the number continues to grow (www.gbif.org/GBIF_org/facility/BIrepfin.pdf).

There are standard practices for treating and preserving new specimens. Plants generally are flattened, dried and glued to reinforcing archival paper. Vertebrate animals may be preserved whole in alcohol or represented only by their skins, bones or shells. Invertebrates like insects are pinned, preserved in alcohol or mounted on microscope slides. In all cases specimens are accompanied by a label

or tag with information about the collector, collection location and date and sometimes by additional information about the habitat or characteristics of the organism. Although the analogy is not perfect, scientists often liken scientific collections to libraries, although books can be reproduced whereas each specimen is unique. Techniques for managing specimen data and book cataloging information are similar; historically specimen data often were maintained in paper catalogues, whereas these data now are being placed in databases, often with Web access (see other articles in this issue).

In recent years the nature of biodiversity specimens has changed, particularly as a result of the application of molecular biological techniques to biodiversity research. Standard specimens now often are accompanied by specimens preserved in ways that protect the organism's DNA, such as freezing, special chemical treatment or rapid desiccation. The DNA itself may be isolated and preserved (usually frozen), and portions may be sequenced and the data deposited in a central database (e.g., GenBank, www.ncbi.nlm.nih.gov/Genbank/index.html). Fungi, protozoans and bacteria frequently are maintained as living cultures, and other living collections like zoos and botanical gardens are sometimes managing their collections as consisting of biodiversity specimens. Frequently a single individual may be represented by several specimens, e.g., a typical museum specimen, frozen tissue and several sequenced DNA regions, necessitating record keeping that retains the connection among the specimens and the associated data, often all in different locations.

Even more significant are changes in the types of data being gathered. In the past, the focus was principally on gathering representative specimens, whereas now the focus often is on the whole ecosystem. Inventories are receiving renewed attention, but they are increasingly quantitative and are often driven by societal needs, such as ecological classification for land use decision-making (including rapid assessments for conservation) and ecosystem health monitoring. Because taking specimens is costly in terms of time, material and storage space and generally requires destructive sampling, there has been an increase in the use of observational data only. Without specimens that can be used to verify identifications, obvious quality assurance issues arise.

There has also been a resurgence in data gathering by trained amateur "citizen scientists" as the demand for inventories outstrips the number of professional scientists available (see, for example, the Illinois Department of Natural Resources EcoWatch Program; dnr.state.il.us/orep/ecowatch/index.htm). A particularly ambitious example of enlisting amateurs is eBird (www.ebird.org), which aims to utilize the popularity of bird watching to compile an extensive bird census database for North America. Quality assurance needs that arise with such data can be addressed in various ways. Developing methods for delivering identification tools to the field, possibly augmented by real-time connections with professional scientists in the laboratory, can improve quality

assurance. Integration of quality control tools directly into the data gathering process offers still further improvement. For example, eBird responds with specific identification queries when rare, difficult-to-identify or out-of-range species are entered into the database. Field data entry using handheld computers linked through satellites to remote databases could further increase data gathering efficiency for both amateurs and professionals and could even be used to generate labels for specimens as they are collected.

In addition to using observational data on organisms, biodiversity studies are increasingly integrating other forms of data. Locations are now being determined using the Global Positioning System (GPS), providing high precision and accuracy. This facilitates integration with remote sensing, mapping and other spatial data through Geographic Information Systems (GIS). Increasingly, quantitative environmental data are also being collected. The United States Longterm Ecological Research (LTER) Network (lternet.edu), which includes 24 sites in different ecosystems, has been collecting such data for about 20 years. These efforts would expand tremendously with the proposed U.S. National Science Foundation National Ecological Observatory Network (NEON). It is estimated that the NEON, when completed, would generate more than 20 million observations daily (www.sdsc.edu/NEON). The demand for efficient data gathering and management will necessitate the participation of information scientists in the planning and execution of the NEON and similar efforts worldwide.

The history of biodiversity data gathering thus shows a significant shift in practice. What was once the work of individuals or small groups working largely in isolation has grown to involve multidisciplinary collaborative projects involving large numbers of people and volumes of data. Field sites are becoming increasingly connected to each other and to scientific institutions worldwide through modern telecommunication links.

Neuroscience

Such significant shifts in scientific practice are equally visible in neuroscience. The application of networking and computational techniques is leading to new forms of collaboration and amalgamation of data. A hundred years ago, data were collected from medical patients and published in the form of case studies. Other data were collected using cadavers' brains or through in vitro animal experiments in which researchers would introduce brain lesions and record the results. (Star, 1989. *Regions of the Mind: Brain research and the quest for scientific certainty*. Stanford, CA: Stanford University Press) Data content included notations on symptoms, treatment information and post-treatment functionality, while data consisted of text and graphics.

Data gathered on the brain during the last several decades has shown significant change, mainly because of changing tools and techniques. Microscopes and electrophysiology sys-

tems were improved and refined. The former allow collection of greater amounts of anatomical data for smaller and increasingly specific parts of the brain. The latter allow electrophysiologists to record multiple trials of electronic “firings” of neuron cells in response to some stimuli, rendering numeric time-series data. Together with scientific developments, these technological innovations have reinvigorated neuroscience, and this is evident in the vast amount of published research. In 1997, the *Journal of Neuroscience* literally doubled its production rate, expanding to two issues per month; in 2001, ISI included 198 titles in its neurosciences journal group; and in 2002-2003, at least two new journals specific to neuroinformatics began publication.

This indication of growth has not been limited to publishing, but is apparent in frequent modification of methods and the increase in the amount of data resulting from the ability to observe and analyze smaller and smaller parts of the nervous system. Work in genetics moved the field toward molecular biology, and electrophysiology and scanning techniques have improved several fold. Neuroscience is a large and diverse field, and there is great variability in research incorporating anatomy, neurophysiology, molecular biology, chemistry and computational modeling. Neuroscience data are gathered in several forms, including numeric, textual, image, graphic and time series. Behavioral and cognitive constructs studied include attention, memory, learning, perception, emotion and language. Research on brain function produces data collected across a number of dimensions, including

- molecular to synaptic, and cellular to systemic;
- single organisms to populations;
- normal through various states of disease;
- birth through old age; and
- across species.

It is understandable that neuroscience investigations often produce large sets of data, but it is easy to overlook the fact that only part of such results may ever be analyzed and reported in a paper. Historically, this data has been collected and stored locally, and there was no tradition of depositing data for future or external use. There is little value in legacy data, which results in years of neuroscience data being disregarded or even lost over time. It should be noted that there is also undeposited data in the biodiversity field; there are many collections “orphaned” as scientists retire. The need for strategies to archive and preserve data is an opportunity for LIS practitioners to engage with scientists to find solutions. It will be particularly important to develop guidelines about what data is important to save, as there is no consensus about this approach (Chicurel, 2000). Research results and new claims continue to be published in traditional ways, and there is grow-

For Further Reading

- Ambrose, S. (1996). *Undaunted courage: Meriwether Lewis, Thomas Jefferson, and the opening of the American West*. Simon & Schuster.
- Chicurel, M. (2000). Databasing the brain. *Nature*, 406 (6798), 822-825. *ebird*. Available at www.ebird.com.
- fMRI Data Center*. Available at www.fmridc.org.
- Human Brain Project* website at www.nimh.nih.gov/neuroinformatics/index.cfm.
- Lewis & Clark Herbarium* website at www.acnatsci.org/museum/lewisclark/l&c_herbarium.html.
- Pachura, C.M. & Martin, J.B. (Eds.). (1991). *Mapping the brain and its functions: Integrating enabling technologies into neuroscience research*. Washington, D.C.: National Academy Press.
- SenseLab* at <http://senselab.med.yale.edu/senselab/>.
- Star, S. L. (1989). *Regions of the mind: Brain research and the quest for scientific certainty*. Stanford, CA: Stanford University Press.
- OECD Megascience Forum Working Group on Biological Informatics. (1999). *Final report of the OECD Megascience Forum Working Group on Biological Informatics* at www.gbif.org/GBIF_org/facility/Blrepfin.pdf.
- Valencia, A. (2002). Search and retrieve. *EMBO Reports*, 3 (5), 396 – 400.
- Working group reports on the *National Ecological Observatory Network* at www.sdsc.edu/NEON.

ing concern about the loss of biological data going forward (Valencia, 2002).

The emergence of *neuroinformatics* in the 1990s has created new possibilities for discovery, but with this there has also been a shift in the public focus, funding and expectations of scientists, and there has been concomitant increase and modification of data collection. The Human Brain Project (HBP) (www.nimh.nih.gov/neuroinformatics/index.cfm) in the United States was started in 1993, following the report of the Institute of Medicine on the need for a coherent agenda for brain research (Pachura and Martin, 1991). From the outset, one of the goals of HBP has been to combine informatics research with neuroscience research. As noted above, biologists studying biodiversity are recognizing the benefits of bringing different and disparate data together; this is beginning in neuroscience. Today HBP supports research projects to federate data queries, refine viewing of cell structures and develop imaging databases that will support meta-analysis and other future use. All of these projects work to bring raw data and supporting materials (such as annotations) together. Other studies have focused on developing systems that store and link textual and image data or computational modeling with neuronal data. (See for example, the *fMRI Data Center* at Dartmouth, www.fmridc.org; the *SenseLab* at Yale, <http://senselab.med.yale.edu/senselab/>)

Shared Problems

The idea of sharing or making public one's data is central to biodiversity informatics and neuroinformatics, and this conflicts with traditional scientific practices. Scientists have voiced a variety of concerns about sharing data; two are widely recognized. Scientists and human subjects (patients or tissue donors), corporations and political entities all need to be assured of the reliability and security of data being deposited in databases and repositories. Neuroinformatics projects that serve as databanks or repositories implement a variety of procedures to improve data quality. The most common tactic is to only accept data from studies that have gone through a peer-review process of a submission to a journal. Another approach is required submission of a set of augmentative information about the experiment or data so that others can judge its value with respect to later use. This information might include, for example, settings on an imaging scanner and the paradigm used for the experimental design. There is some debate among

biologists and other researchers about what parts of this material constitute metadata. However, gleaning and managing the significant information related to shared data is important and necessary for both biodiversity and neuroscience.

Although biodiversity studies and neuroscience differ significantly in scope and focus, both will depend on information science to provide needed expertise if they are to effectively address scientific and societal needs. People doing many different types of work utilize biological specimens. This variety is not so much the case in neuroscience. For obvious reasons amateur "citizen scientists" are not involved in neuroscience. Yet even though the heterogeneity of biodiversity data is qualitatively different from neuroscience data, there are many ways that these fields would similarly benefit from the traditional knowledge of library and information science. Information science researchers and practitioners can bring expertise in data visualization and retrieval techniques, records management, quality assurance and usability.

The **BULLETIN OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY** is a **BIMONTHLY PUBLICATION** that serves as the newsletter of the Society. It publishes short articles on a **BROAD RANGE OF TOPICS** of current concern to **ASIST MEMBERS**, focusing particularly on material of interest to practitioners. Readers are **ENCOURAGED TO SUGGEST** topics of interest or alert the Editor of suitable material that may have been presented at **ASIST-sponsored events or elsewhere**. In addition, authors are **ENCOURAGED TO SUBMIT** articles on topics such as **CURRENT PRACTICE, PUBLIC POLICY, LEGISLATION, STANDARDS, PILOT PROJECTS, STATE-OF-THE ART REVIEWS or OVERVIEWS OF EVOLVING TECHNOLOGY AND ITS IMPACT**. Articles informing the membership about various developments within **ASIST** are very welcome, as are articles reporting on **ACTIVITIES OUTSIDE THE UNITED STATES**. The *Bulletin* encourages original articles, but will consider **TIMELY MATERIAL** that has been presented or published elsewhere. Articles are posted in full on the **ASIS Web Site** at <http://www.asis.org/Bulletin/index.html>

Authors interested in developing material for a focused issue are urged to contact the Editor directly.

Authors are encouraged to discuss article ideas with the Editor if there are questions about suitability or relevance.

Irene L. Travis, Editor
Bulletin of the American Society for Information Science and Technology
 1320 Fenwick Lane,
 Silver Spring, MD 20910
 (301) 495-0900
 Bulletin@asis.org

BULLETIN
 of the American Society for Information Science and Technology