

# Toward Integrative Science: Organizing Biodiversity and Neuroscience Data

by Melissa H. Cragin

---

*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.*

---

The 1999 OECD report on biological informatics brought the biodiversity and neuroscience fields together into a single framework to describe needs and barriers to global information access. These two fields share a common orientation, focused on finding and identifying objects, naming new organisms and structures, and discovering and describing the relationships between them. The need for integrative science in the biodiversity and neuroscience fields has real consequences, and many important questions will only be addressed by integrating data from many and varied sources. Often such integration will mean connecting data derived from different levels of analysis, different species, even different fields. This paper presents some of the challenges researchers face in bringing data sets together for retrieval and use, including the creation of data collections, collection interoperability, and access and use.

Both fields are entering a stage of development where new discoveries will be made by learning how systems and subsystems behave. Without an integrative approach to biodiversity data, political and economic policy decisions are made with very little information about the big picture. That is, understanding biodiversity on a global scale will require research beyond the organismal level. It will require an understanding of the land itself, as well as weather and other phenomena. Data collected for very different purposes, like rainfall data, will need to be included in retrieval systems for analysis along with species data. Similarly, in neuroscience, advances are made daily that support our understanding of individual cells and subsystems of the brain, but developing a complete picture of the brain will require methods that can sup-

port evaluation of processes which are themselves integrative. To understand the processes of normal nervous system function and disease states, scientists will need to synthesize findings from different levels of study and determine their interactions and relationships.

There are also key differences between biodiversity and neuroscience, such as the structures of the fields, the management of data and the traditions of interaction with other (or sub) disciplines. Biologists and systematists, for instance, recognize the need for communication and collaboration with other fields such as ecology. They have come to value wide access to biodiversity databases that will connect with data sets created by a variety of owners. Neuroscientists, however, do not have a tradition of sharing raw data (Chicurel, 2000; Koslow, 2000); there are no projects to digitize legacy data for future use. Also, although neuroscience may be more interdisciplinary in practice than the field of biodiversity, it is a very “fractured” science in which subspecialties have developed along several dimensions (Toga, 2002). Specialties such as neuroimaging, electrophysiology, morphometrics and computational modeling each shape practices resulting in an array of preferred instrumentation and methods.

In order to take advantage of informatics developments scientists will need to change some of their research practices. The most prominent concerns identified in the literature of both fields are about data quality and trust, data sharing, communications with external groups or disciplines and rights management. Just as important but less discussed are the ways that researchers think about future uses of their data, including the following:

- How does their data relate to research outside of their specialty?
- Are there possible uses for their data in different kinds of studies?
- What information will be necessary to accompany a data set so that it can be re-used?

### Gathering Data Together into Collections

Successful integrative science will require more open scientific practices in which scientists are cognizant of possible use of their data in the future, and institutions value and reward activities beyond traditional publication (Amari, et al, 2002). For the scientists such practices will include documenting the processes of data production and submitting that data and documentation to a repository or building a new shared database. For institutions this will require recognizing these new kinds of contributions in tenure and promotion processes. For long-term use databases and repositories will need curatorial management and archiving expertise to preserve the contents. Considerations of sustainability should be a mandatory part of these large-scale projects, and this will require addressing issues like funding, control and access. First, however, there needs to be consideration of what data should be included in shared databases and repositories.

**Legacy Data.** Legacy data are those that result from research conducted in the past that are not yet stored digitally. In biodiversity these data include museum data, conservation data and data found in individual labs. Some biodiversity data are simple and relatively homogenous. Such data generally consist of label data from biological specimens, including the scientific name of the specimen, where and when the specimen was found and who found it. However, as noted elsewhere, knowledge about the species that inhabit the planet will depend on the integration of data from other fields. If label data from many collections over a 100-year period are pooled and combined with forest fire data, it would be possible to evaluate the role of fire in species distribution, which could help set burn and forest clearing policy.

**Integrating Scales of Analysis.** A primary goal of biological informatics is to shift the focus from research on discrete

pieces of a biological puzzle to whole sections of the puzzle. Sometimes this work involves the integration and analysis of data at different scales. Some of the most vivid examples of this can be seen in the work conducted at the National Center for Microscopy and Imaging Research (NCMIR) where informatics researchers are designing tools for scientists to literally see microscopic structures in degrees. Scientists use these tools to locate and render images of very small parts that actually vary in size, from spine to dendrite, synapse to neuron to cell cluster. Scientists are creating images of materials that scale with a factor of ten or more. The complexity of merging such different units of analysis is uniquely challenging, and will require new visualization and classification tools.

**Developing Collections of Tools.** The collection of analysis and visualization tools is already an important aspect of biological data systems. For access to the tools at the NCMIR, scientists may either come to the center or utilize telecomputing. This serves the research interests of both the scientists, who may not otherwise have access to these tools, and the center that is fostering collaborative research. Neuroscience has traditionally been a local activity – data systems and tools have been developed by or for the specific labs in which they are used. This practice continues, though now tools are also developed by those building the common databases and repositories with the intention that anyone can use the tools when they access the data. However,

some scientists still want to be able to use their own locally developed tools on this shared data. To address this desire, some initiatives such as the Biomedical Informatics Research Network are collecting locally developed tools and making them centrally available to users, while protecting the “ownership” rights of the developers. Still, proprietary issues and concerns about standardization persist in the field, and these issues point to the kinds of theoretical problems described by Bowker (2001a, 2001b) on the integration of biodiversity data.

**Accumulating and Mobilizing Literatures.** Neuroscience is a very broad field, with subdisciplines spanning anatomy, molecular biology, neurochemistry, electrophysiology, neuroimaging and behavior. As noted, research has become highly specialized, and increasing fragmentation results in a large

“The Human Brain Project will shape, into the next century, an important frontier of science and technology. . . . The long-term goal of this initiative is to integrate brain and behavioral research with informatics research and development, giving scientists, students and clinicians intelligent access to the full range of information about the brain. A newly integrated view of the brain will facilitate the generation of hypotheses that address broad issues that are significant to the overall understanding of human health and disease.” – Koslow, S.H., & Huerta, M.F. (Eds.). (1997). *Neuroinformatics: An overview of the Human Brain Project*. Mahwah, NJ: Lawrence Erlbaum Associates, p. 9-10.

## Although computer storage devices continue to increase in capacity, they will not be able to hold everything. Further, disparate sets of data require different kinds of organization for management and querying.

scatter factor. Researchers acknowledge that they have difficulty “keeping up,” even in their own areas of expertise. An early goal of the Human Brain Project was the development of an atlas of the human brain “to be used as a framework to index and link to databases for other types of brain data, including those not necessarily captured by neuroimaging (e.g., histological data, models of function of particular brain structures, directories of scientists studying certain brain structures, and bibliographic data)” (Huerta & Koslow, 1996, p. S 5). Although such a unified atlas has not been developed, there are several atlas projects underway that will link to other kinds of data – click on a particular label and launch a query.

### Collection Interoperability – Databases and Repositories

There are a variety of technological and social barriers to the creation of large-scale interoperable systems. Successful implementation of these systems depends upon social agreements about (discipline-specific or specialized) standards for conducting (and thus describing) research activities, standards of validity and controlled vocabularies. Groups in most scientific communities are grappling with these issues and the processes of enacting their subsequent arrangements. Understanding how different disciplines communicate, interact with information resources and share data sets will help systems and metadata developers meet users’ needs. However, it is likely that the following technological challenges may be less difficult to solve than many of the present social barriers.

**Volume.** The trend in the mid-90s was to try to build centralized databases for accumulation of research data. However, tools and methods have facilitated increasingly larger sets of data – geospatial data and brain imaging data can easily reach the terabyte level. Informaticists, computer scientists and domain experts have realized that centralized databases were not a viable approach to meet these scientific needs. Although computer storage devices continue to increase in capacity, they will not be able to hold everything. Further, disparate sets of data require different kinds of organization for management and querying. Distributed, federated systems will support long-term maintenance and allow faster and more flexible information retrieval (Jones, et al, 2001; Van Horn, et al, 2001).

**Heterogeneity and Complexity.** Biological data in its entirety represents concepts and structures from multiple domains that

study the earth and its creatures along several dimensions. In addition, biodiversity researchers will need to use data from multiple new sources to develop future interpretations of species data. These disparate data types include data on land formations, climate changes, ecologies, animal and plant physiology and anatomy, biological systems, molecular and chemical biology, and behavior. Much of this data, collected by researchers and scientists in other fields such as climatology, was recorded for different purposes than to meet the analytic needs of biologists. Such variety presents a range of integration and re-use problems. By contrast in neuroscience, there is a singular purpose to the research – to make discoveries about the structure and function of the central nervous system, particularly the brain. However, the range of disciplines and research methods produce many forms of data, including numeric, textual, image and time-series. The resulting heterogeneity is not trivial, and this presents additional challenges to informatics researchers.

**Item-Level Metadata.** In the context of biological informatics, metadata is an interoperability issue. It is common to require metadata be submitted to databases along with research results, where it is tied to the validity and value of the scientific data itself. To preserve legacy data and increase the value of findings from current research, describing the experiments and data from that research is a necessity. The fMRI Data Center at Dartmouth, for example, has established four areas of metadata that depositors must include – protocol (about technique), human subject information, scan session information and experimental protocol (Van Horn, et al, 2001). This is a specific area where library and information scientists can contribute to the development of biological data collections, as our field is actively engaged in creating and testing solutions for metadata systems.

### Access and Use

There is now consensus within biodiversity and neuroscience informatics fields that a distributed architecture with federation will best support the scale and complexity of the data. This approach will address the frequent concern about data control by keeping the location and control of these databases at their local institutions and countries.

**Nomenclature.** There are problems of nomenclature that will need to be solved. In biodiversity there are examples of multiple naming, where “new” discoveries of previously named species are given different names. In addition, millions of

species remain unnamed. To name these new species we must sift through the tens of millions of named species to insure they were not previously named. Electronic systems support annotations and current rules for addressing this problem by making the different specimen types distinguishable. However, more will need to be done to support ongoing disambiguation, and there is a general lack of funds to support the basic taxonomy. Better global access to data sources like electronic flora and fauna will facilitate the information work that is required to verify either novelty or accurate identification of a specimen.

In neuroscience, there are at least two distinct problems with nomenclature. First, many regions of the brain do not have inherently or anatomically distinct boundaries. What Dr. Smith calls cell cluster “A,” Dr. Jones includes in his identification of cell cluster “T.” The second type of problem is related to synonyms; there may be multiple terms representing a single action or concept. Current IR systems are not very good at returning results across the range of terms. Calls for standardization have been met with resistance, as scientists are hesitant to give up the meaning that attaches to their particular terms. Library and information scientists specialize in these sorts of problems and could contribute here by helping to develop powerful thesauri and new retrieval tools.

**Federating Data and Queries.** Building collection-level metadata repositories will also help to support global research. These repositories will allow queries to be returned faster, as they will first find the collections from which the query might be answered. “Query-time federation” is being developed using at least two models – one that steers queries through a central metadata repository which then directs the query to (appropriate) databases for retrieval and the other using a mediating protocol such as Z39.50 where the client queries each of the repositories when the user executes a query.

**Quality Assurance.** Scientists and researchers want to have some trust in the reliability and specific application of their data and tools. Informatics projects are testing ways to support data quality, using variations of peer-review, annotation systems and cross-referencing, data weighting and others (Amari, et al, 2002; Chicurel, 2000). These systems have not yet seen systematic measurement, so it is too early to know which approaches will provide the best way to meet the needs of the various user groups. In biodiversity, the very act of digitizing museum specimens opens up the possibility of data validation. For example, scientists and others worldwide will be able to evaluate the validity of the name of the specimen using its location information and general knowledge of taxonomic revisions.

## Conclusions

There are economical, political, social and technological challenges to bringing biological data together in a way that will permit new kinds of integrative analysis (see other articles

in this issue). Technological and intellectual challenges include the lack of standards, developing and testing novel integration and retrieval techniques, and building thesauri and vocabulary systems. Information science can help solve the problems of managing and analyzing growing amounts of heterogeneous data. The movement in biological sciences toward global and systems level knowledge will require the sharing and integration of disparate data and tools that will support testing new kinds of questions. Finally, expanded study is needed to refine systems development, to illuminate “triggers” of collaboration and describe the implications for scientific practice and scholarly/scientific communication.

## For Further Reading

- Amari, S-I., Beltrame, F., Bjaalie, J.G., et al. (The OECD Neuroinformatics Working Group). (2002). Neuroinformatics: The integration of shared databases and tools towards integrative neuroscience. *Journal of Integrative Neuroscience*, 1 (2), 117-128.
- Biomedical Informatics Research Network (BIRN). Available at <http://www.nbirn.net/>
- Bowker, G. (2001a). Biodiversity, datadiversity. *Social Studies of Science*, 30 (5), 643-683.
- Bowker, G. (2001b). Mapping Biodiversity. *International Journal of Geographical Information Science*, 14 (8), 739-754.
- Chicurel, M. (2000). Databasing the brain. *Nature*, 406, (6798), 822-825.
- Huerta, M. F., & Koslow, S. H. (1996). Neuroinformatics: Opportunities across disciplinary and national borders. *Neuroimage*, 4 (3), S4-S6.
- Jones, M. B., Berkley, C., Bojilva, J., & Schildhauer, M. (2001). Managing scientific metadata. *IEEE Internet Computing*, 5 (5), 59-68.
- Koslow, S. K. (2000). Should the neuroscience community make a paradigm shift to sharing primary data? *Nature Neuroscience*, 3 (9), 863-865.
- OECD Megascience Forum Working Group on Biological Informatics. (1999). *Final report of the OECD Megascience Forum Working Group on Biological Informatics*. Available at [www.gbif.org/GBIF\\_org/facility/Blrepfin.pdf](http://www.gbif.org/GBIF_org/facility/Blrepfin.pdf).
- Toga, A. W. (2002). Neuroimaging databases: The good, the bad, and the ugly. *Nature Reviews Neuroscience*, 3 (4), 302-309.
- Van Horn, J. D., Grethe, J.S., Kostelec, P., Woodward, J. B., Aslam, J. A., Rus, D., Rockmore, D., & Gazzaniga, M. (2001). The Functional Magnetic Resonance Imaging Data Center (fMRIDC): The challenges and rewards of large-scale databasing of neuroimaging studies. *Philosophical Transactions of the Royal Society of London: Biological Sciences*, 356 (1412), 1323-1339.