Information Management Challenges to Integrated Inventory and Monitoring of Forest Ecosystem Resources[1]

William K. Michener[2]

Abstract—Many complex research questions will guide science in the 21st century including issues related to sustainable production, protection of plant and animal diversity, and global climate change. A federated information infrastructure that can provide seamless access to shared data and information resources is required to address these issues. Impediments to developing this federated information infrastructure include technical (e.g., communication infrastructure, database interoperability, data archives), semantic (e.g., metadata, methods standardization), and social (e.g., scientific reward structure, cross-disciplinary communication) challenges. Although many of the challenges require long-term fixes, several proactive steps can be taken now to facilitate integrated research and monitoring of forest ecosystem resources.

Environmental scientists have become increasingly interested in issues related to sustainable production; climate and land use change; air, water, and soil pollution; decreases in plant and animal diversity; and human dimensions of global change. To address these issues, scientists and resource managers are increasingly working at broader spatial and longer temporal scales. Attempts to scale research to the region, continent, and globe require unprecedented collaboration among scientists, data sharing across borders, and ready access to high quality, well-documented data that have been preserved in data archives.

Data and information management represents a process that starts with project design and extends beyond the data analysis and publication phases. For example, information management within an organizational context includes design of paper and digital data entry forms, quality assurance and quality control, data processing (e.g., subsetting, merging), metadata development, and submission of data and metadata to a data center or data archive. Generally, a successful research program at an institution depends on a high quality information management program (see Michener et al. 1994, 1998).

Good institutional information management does not guarantee research success, particularly when the scope of inquiry is expanded to regional and global scales, and data from many institutions are required to address specific questions. Such attempts to broaden research efforts often fail because the data pertaining to environmental resources are incompatible, uncoordinated, stored in isolated locations, inadequately protected, and poorly documented. In essence, the requisite data are inaccessible or inadequate. In recognition of this persistent problem, a recent President's Committee of Advisors on Science and Technology proposed development of a federated information infrastructure whereby terabytes of data from many different sources (satellite, field, and laboratory) can be efficiently searched, data can be readily compiled in new ways for analysis and synthesis, and the resulting information can be presented in understandable and useful formats (PCAST 1998; also see Robbins 1996).

Numerous technical, semantic, and social challenges are inherent in developing the federated information infrastructure that is envisioned (Michener et al. 1994, Robbins 1996, PCAST 1998). In the following discussion, I outline impediments to a federated information infrastructure, many of which will require long-term funding, research, and infrastructure improvements. However, other actions can now be taken by various organizations to improve their information management capabilities and pave the way to the more responsive science and resource management needed for the 21st century.

Information Management Challenges

There are at least three different categories of impediments to achieving a true federated information infrastructure that can meet the information needs of science in the 21st century. These include technical, semantic, and social challenges.

Technical Challenges

Technical challenges include the need for an improved communication infrastructure (i.e., Internet-2) that can handle massive bandwidth requirements and advanced network architectures. Database interoperability must be significantly improved. New and expanded data archives will be necessary to provide both secure storage and ready access to environmental data.

Many of the most difficult challenges that lie ahead relate to translating data into information and, ultimately, knowledge. Data, for instance, consist entirely of characters and numbers that have little or no intrinsic meaning. On the other hand, information is a much higher level representation of data where the data have been given form or character, and confer meaning. Knowledge is the understanding that is gained through discovery, perception, and erudition of

[1]Paper presented at the North American Science Symposium: Toward a Unified Framework for Inventorying and Monitoring Forest Ecosystem Resources, Guadalajara, Mexico, November 1-6, 1998.

[2]William Michener is Associate Scientist, Joseph W. Jones Ecological Research Center, Route 2, Box 2324, Newton, Georgia 31770, USA. Telephone: (912) 734-4706. Fax: (912) 734-4707. E-mail address: wmichene@jonesctr.org

information. From a scientific standpoint, the true value of data is directly related to our ability to extract higher level understanding from those data.

Knowledge about our environment entails the synthesis of data from many sources and typically requires that a human being has to acquire, manage, manipulate, correlate, analyze, and synthesize data from individual data sets, one at a time. The continuance of long-term monitoring programs coupled with significant improvements in sensors and data acquisition have led to the current situation where many organizations now have an excess of data. Exponential increases in the size of data holdings coupled with the recent development of new environmental remote sensing technology (e.g., multispectral data at 1-3m^2 spatial resolution) requires that we develop new approaches to exploit these massive data sets. Specifically, we need tools to analyze and synthesize data quickly and translate those data into useful information that can guide decision-making, policy formulation, and future research.

Critical technological challenges include the need to develop:

- Extraction and analytical tools for correlating, manipulating, analyzing and presenting distributed information (e.g., new analytical (statistical and modeling) techniques that work with multidimensional, large-volume data).
- New quality assurance methods that "correct" data errors with minimal human intervention.
- Metadata encoding routines to facilitate data mining of these massive data sets.
- Algorithms for analysis, change detection, and visualization that scale to large, multi-temporal, and multi-thematic databases.

Semantic Challenges

Semantic challenges encompass those factors that lead to difficulties in understanding and interpreting data. Environmental data are particularly complex. Forest ecosystem data encompass the complexity of millions of different organisms and hundreds to thousands of different communities and ecosystems. The data are collected by different countries, agencies, industries, academic institutions, and individuals, all of which have different needs, views, requirements, and skills. Furthermore, the data vary substantially in scale, precision, accuracy, type (text, measurements, images, sound, video), and volume (kilobytes to terabytes). Consequently, even within a single institution, data pertaining to one environmental parameter often cannot be compared with data on other parameters because of different data structures, scales of measurement, region of coverage, times data were collected, and so on. Problems associated with different data collection protocols, data storage mechanisms, and comprehensiveness of data documentation are compounded when more than one institution is involved.

Two specific actions that can minimize semantic conflicts include the development and adoption of metadata content standards, as well as the standardization of data collection protocols where possible. Metadata provide critical information for expanding the scales at which ecologists work. For example, field validation data from multiple sites are frequently used to calibrate (or, in some cases, are merged with) remotely sensed data, thereby expanding the spatial domain from the site to broader scales. Cross-site comparative studies depend heavily upon the availability of sufficient metadata. For cross-site comparisons, it is especially important that both methods and instrumentation calibration and inter-calibration (measurements of similar parameters by different methods or instruments) be well documented to confirm data integrity, proper use of experimental methods, and data acquisition.

Much of the *post hoc* effort that is devoted to managing (manipulating), merging, and analyzing data for parameters that were collected under different protocols can be reduced or eliminated when standard methods are employed *a priori*. Furthermore, development and adoption of standard data collection and management protocols also reduces the amount of time and effort expended in developing metadata.

Social Challenges

Research and resource management for the 21st century will require unprecedented collaboration among scientists from many disciplines, as well as data sharing across departmental, agency, academic, and national borders. Success will depend on the extent to which we alter existing scientific reward structures. Data sharing and collaboration are facilitated when the stakeholders perceive that there are real benefits in doing so (Porter and Callahan 1994). Thus, if the delivery of useful data products is part of an organization's objective, then those contributing to the development of the product (i.e., database) deserve credit for doing so. In essence, databases should be viewed as being synonymous to a publication and should be considered in personnel review and promotion procedures.

The lack of effective cross-disciplinary communication, data sharing, and collaboration often impedes attempts to broaden the scope of our scientific efforts. Strategies for success include: (1) taking that first step to initiating a dialogue; 90% of success is simply showing up; (2) build upon existing successful partnerships (i.e. past successes); (3) plan early for collaborative efforts; and (4) provide incentives and the reward structure for participation in cross-disciplinary ventures. Communication is the key to resolving conflicts among participants with different training, vocabularies, and world views.

Implications for Inventory and Monitoring

Research challenges include striking a rational balance between economic feasibility and the data scale(s) and volume that are optimally required to meet scientific and monitoring objectives. High quality, well documented, securely preserved, and accessible data are essential for addressing long-term and broad-scale environmental problems. Access to high quality data requires a strong commitment to implementation of effective information management procedures. The absence of such procedures impairs our ability to use data over long periods. For instance, the loss of information content associated with data through

the degradation of the raw data or the metadata is unavoidable and has been referred to as "data entropy" (Michener et al. 1997). Adherence to recommended data management practices, especially the development of comprehensive metadata and the submission of both data and metadata to data archives greatly slows the progression of data entropy.

In this discussion, I have listed many of the challenges to developing the federated information infrastructure required for science and management in the 21st century. Solutions to many of the problems will require substantial infusion of money, personnel, creative thought, and technology by businesses, research and resource management organizations, and nations. However, there are at least seven steps (habits) related to information management that can be taken now to facilitate highly effective integrated research and monitoring of forest ecosystem resources:

1. Allocate a reasonable percentage of research funding for long-term management of data and information generated by the research. In most organizations, data management is seriously under-funded, resulting in data losses and delays in translating data to information.
2. Develop and adhere to data and metadata standards and best use protocols.
3. Provide funding for data rejuvenation (e.g., adding Global Positioning System fixes, i.e., latitude/longitude, to field sites) and rescue (e.g., convert paper records to digital format) to halt further data entropy.
4. Routinely evaluate data utility, research objectives, and management needs, and reestablish priorities. Use this information to revise sampling programs (e.g., reduce effort in certain areas, add new parameters) and to streamline data capture.
5. Coordinate software and systems development and purchases with other agencies or departments to eliminate duplication of effort and reduce expenditures (i.e., take advantage of economies of scale).
6. Cooperate with other agencies, scientists, and the private sector to establish and adopt data and metadata standards, authority files, and thesauruses for data.
7. Establish synthetic research as a top priority.

Literature Cited

Michener, W.K., J.W. Brunt, and S.G. Stafford. 1994. Environmental Information Management and Analysis: Ecosystem to Global Scales. Taylor and Francis, Ltd., London, England.

Michener, W.K., J.W. Brunt, J. Helly, T.B. Kirchner, and S.G. Stafford. 1997. Non-geospatial metadata for the ecological sciences. Ecological Applications 7:330-342.

Michener, W.K., J.H. Porter, and S.G. Stafford (eds.). 1998. Data and Information Management in the Ecological Sciences: A Resource Guide. University of New Mexico, Albuquerque, NM. {Available through http://www.lternet.edu/ecoinformatics/guide/frame.htm}

(PCAST) President's Committee of Advisors on Science and Technology. 1998. Teaming with Life: Investing in Science to Understand and Use America's Living Capital. Office of Science and Technology Policy, Washington, DC. (Available through http://www.whitehouse.gov/)

Porter, J.H. and J.T. Callahan. 1994. Circumventing a dilemma: Historical approaches to data sharing in ecological research. In: Environmental Information Management and Analysis: Ecosystem to Global Scales (eds W.K. Michener, J.W. Brunt, and S.G. Stafford), pp. 193-203. Taylor and Francis, Ltd., London, England.

Robbins, R.J. 1996. Bioinformatics: Essential infrastructure for global biology. Journal of Computational Biology 3:465-478.