The New Bioinformatics: Integrating Ecological Data from the Gene to the Biosphere

Matthew B. Jones,[1] Mark P. Schildhauer,[1] O.J. Reichman,[1] and Shawn Bowers[2]

[1]National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara, California 93101; email: jones@nceas.ucsb.edu, schild@nceas.ucsb.edu, reichman@nceas.ucsb.edu

[2]Genome Center, University of California, Davis, California 95616; email: sbowers@ucdavis.edu

Annu. Rev. Ecol. Evol. Syst. 2006. 37:519–44

First published online as a Review in Advance on August 14, 2006

The *Annual Review of Ecology, Evolution, and Systematics* is online at http://ecolsys.annualreviews.org

This article's doi: 10.1146/annurev.ecolsys.37.091305.110031

1543-592X/06/1201-0519$20.00

Key Words

ecoinformatics, data integration, data sharing, metadata, ontology, scientific workflows, semantics

Abstract

Bioinformatics, the application of computational tools to the management and analysis of biological data, has stimulated rapid research advances in genomics through the development of data archives such as GenBank, and similar progress is just beginning within ecology. One reason for the belated adoption of informatics approaches in ecology is the breadth of ecologically pertinent data (from genes to the biosphere) and its highly heterogeneous nature. The variety of formats, logical structures, and sampling methods in ecology create significant challenges. Cultural barriers further impede progress, especially for the creation and adoption of data standards. Here we describe informatics frameworks for ecology, from subject-specific data warehouses, to generic data collections that use detailed metadata descriptions and formal ontologies to catalog and cross-reference information. Combining these approaches with automated data integration techniques and scientific workflow systems will maximize the value of data and open new frontiers for research in ecology.

1. INTRODUCTION

Bioinformatics: the use of computational and statistical techniques to more effectively manage and analyze biological data

Ecoinformatics: a field of research and development focused on the interface between ecology, computer science, and information technology

In 2003, *Science* and *Nature* simultaneously published cover stories about the demise of gorillas in Africa (Kaiser 2003, Whitfield 2003). The title of the story in *Science* ("Ebola, Hunting Push Ape Populations to the Brink of Extinction") revealed that the ecological issues were quite broad. Ebola, a pernicious hemorrhagic disease that can spread to humans, is affected by its local environment. Understanding this disease requires knowledge of epidemiology, genetics, and transmission modes, along with their ecological contexts. Hunting pressure engendered by the need for bushmeat relates to the nutritional status of local humans and sociological features of their culture. With regard to the apes, information about their population dynamics (birth and death rates, longevity, social interactions) is needed to take timely, effective action to save the species. Virtually every ecological question, whether this dire or not, requires access to a similarly diverse array of data and information in order to develop robust analyses. Integrating ecologically pertinent data into the chain of information from the gene to the biosphere will significantly enhance our understanding of the natural world and promote wise management strategies for natural resources. In this review, we examine challenges and solutions relative to locating, accessing, integrating, and analyzing data from ecology and allied disciplines.

2. THE NEED FOR A NEW BIOINFORMATICS

Bioinformatics is the application of techniques from computer science and statistics to manage and analyze biological data. The initial focus within bioinformatics has been on tools and analytical techniques that operate on genetic and protein sequence data (many useful databases have emerged in this context, including GenBank; Benson et al. 2005) and there is ongoing discussion about the need to further integrate these resources with higher systems levels such as data describing biological processes at the metabolic level (Thomas & Ganji 2006).

Ecology as a discipline grew out of a natural history tradition with a strong emphasis on observation in the field. By the late nineteenth century, ecology was becoming a more quantitative science with fewer purely descriptive studies. Later studies increasingly moved toward mathematically derived models that focused on assessments of the distribution and abundance of organisms along with related information about the abiotic environment (Real & Brown 1991). Since the 1960s there has been a strong emphasis on experimental manipulation to elucidate causal relationships (e.g., Brown & Munger 1985, Connell 1961, Lubchenco & Real 1991, Paine 1966). Experiments are typically designed to test a particular set of hypotheses and therefore the types of manipulations performed and the formats of data collected vary tremendously across studies. These factors contribute to making ecological data highly heterogeneous.

The most significant challenge in ecological informatics (ecoinformatics) is dealing with the inherent complexity and breadth of data used in ecological studies. Ecological data do not only document entities of interest—such as the numbers of individuals, or sequences of nucleotides. Rather, they frequently contain measurements of processes (e.g., rates of competition, or herbivory) or surrogates for these (extent of shading,

or assessment of leaf damage) that often require specialized expertise to accurately document and interpret. Ecological data also occur in many forms (text, numbers, images, videos), and numerous legacy data that are important for dealing with scientific and environmental issues remain undigitized. These characteristics make the access, interpretation, analysis, and modeling of ecological data especially challenging.

Ecologists have recognized the need for integrated data systems to support cross-disciplinary collaboration to understand the basic ecological principles that govern the biosphere (Green et al. 2005). With the rapid growth of human populations and their impacts, it becomes critically important to better describe and understand natural processes. The increasing demands within ecology for greater access to more types of data emphasize the need for integrated data-management solutions that span biological subdisciplines from the gene to the biosphere.

3. CASE STUDIES IN SYNTHESIS

Although data from the observations and experiments of individual investigators remain at the core of ecological and evolutionary research, their value increases substantially when they are integrated and synthesized to reveal important patterns and to generate broad generalities. Data synthesis allows a broader perspective over time and space, and across many disciplines, than is possible from one or a few studies. Even more important in the long run, synthesis allows data to be used for purposes other than those for which they were originally intended, to address questions that were unknown or unapproachable at the time the data were collected (e.g., Andelman et al. 2004).

Regardless of whether synthetic and integrative research is undertaken in collaborative frameworks (such as is sponsored by various synthesis centers) or by individuals, these efforts will almost certainly depend on access to complementary data that were collected by other individuals or under the auspices of other projects. Once the data needs extend beyond local data collection efforts specifically tailored for a given analysis, efficient access to data will be severely hampered for a variety or reasons, ranging from difficulties owing to sociological and legal reasons (e.g., lack of permission to use data) to technical issues such as data contained within incompatible management systems, or integration challenges resulting from variable spatial and temporal scales of sampling, taxonomic irregularities in the identification of specimens, and idiosyncratic labeling of variables and their units of measurement.

A project comparing the effects of grazers and fire on grasslands in North America (Konza Prairie) and South Africa (Kruger Park; Knapp et al. 2004) provides an excellent example of both the difficulties of synthesizing data from multiple sources and creative solutions for dealing with them. The researchers dealt with incongruous variables that quantified the effects of treatments by using those that were similar, and correlating surrogates for other variables that were dissimilar. Many plant taxa were sampled in one study, whereas only trees and grasses were sampled in the other, so the researchers had to use growth form rather than taxon, per se, to compare data. Plot size and methodology differed so data from each location were transformed into relative abundances. Because the fire and grazing regimes were imposed differently at

the two locations (experimental versus natural), the treatments had to be reduced to ordinal rankings (e.g., high or low fire frequency) to analyze how each affected plant productivity and community-level characteristics. Significant amounts of data had to be discarded because of incompatibilities among variables in the source data sets. By comparing the consolidated measures within a site (e.g., converting actual plant abundance counts to relative abundance, species to growth form, and actual fire frequencies to relative frequencies) the researchers conducted a robust synthetic analysis using the data. Still, the process of integrating the data was arduous and inefficient, and the strength of the analyses was impacted by the inability to incorporate all the available data and the need to use ordinal rankings instead of the original numeric values.

Relational database: the prevalent software method for storing tabular information that uses named 2-dimensional tables

Data model (data schema): how information should be conceptually and logically described to optimize storage, access, and interpretation in some computer application

4. CURRENT METHODS FOR STORING AND ACCESSING ECOLOGICAL DATA

The most common method that scientists currently use to manage ecological data is to enter it in an ad-hoc manner in spreadsheet-based software tools. Spreadsheets are flexible, easy to learn, and allow scientists to quickly enter, review, and get summaries of data in a simple although statistically unreliable manner (McCullough & Wilson 1999). Spreadsheets do not provide the tools to promote good data management practices, however, because they lack sufficient structure to adequately describe and constrain the data. Spreadsheets often contain multiple data tables on a single page, and easily permit intermixing of raw data values with statistical summarizations, and marginal sums and annotations. For the scientist, using a spreadsheet in this way is often convenient over the short term. But unless these usages are well-documented, such informal data management practices cause difficulty for scientists looking back at their own data and inhibit reuse of the data by other scientists that may be unfamiliar with the organization of the spreadsheet. Although having data available in any format is probably more valuable than not having it at all, automated data processing approaches will always be constrained by the unstructured way in which spreadsheets store data.

Scientists seeking a more robust way to store their data frequently learn to use relational database systems such as Microsoft Access or Filemaker Pro. Alternatively, many ecologists store and analyze their data in a statistical package such as SAS (**http://www.sas.com**) or the "R Statistical Package" (**http://www.r-project.org**). In all these cases, a researcher typically models the data beforehand by deciding how to separate them into tables (usually with columns representing variables, and rows equaling observations). The researcher then specifies how these tables, which each contain some distinct conceptual type of information, or entity, can be joined or merged (Brunt 2000, Pascal 2000, Porter 2000). For example, one might join a table containing observations of species abundance on a given day with a table of observations of the temperature taken at that same location through time to create a new table that has information from both. Ecological researchers usually learn these skills on their own, because data modeling and implementation in a database management system (DBMS) are not currently standard parts of a biologist's academic

training, despite their relevance for better understanding how to collect and manage data.

Owing to limitations in desktop DBMSs and the computer operating systems on which they run, it is relatively difficult to share databases and spreadsheets with colleagues. This problem grows with time as the software on which the data depends becomes obsolete and is replaced by newer tools, leaving the older proprietary databases and spreadsheets in an inaccessible state. Moreover, the data in these systems is typically structured to serve the specific needs of the project. The practice of creating project-specific data sets prevents standardization of approaches among otherwise similar studies and ultimately leads to data integration challenges for synthetic analyses.

Data integration: matching up and combining information from different sources, ideally in ways that are meaningful and useful

4.1. Data Warehouses (Vertically Integrated Databases)

One solution to the problem of project-specific databases complicating data integration is to develop vertically integrated databases that store data collected by many different investigators, but all following a common theme. These data systems are often called data warehouses, and usually have a Web-based interface for querying and downloading data. They are thus broadly accessible to individuals, without the need to locally store the data, or install specialized DBMS software. Examples of vertically integrated data warehouses include centralized data archives such as GenBank, VegBank (Jennings et al. 2004), and TreeBase (Morell 1996); others provide network-distributed access to data from a number of compatible servers (e.g., access to biodiversity and specimen collection data in the GBIF portal; Canhos et al. 2004). These databases typically are more complex than desktop databases because they attempt to reconcile the differences in the data models among existing independent research projects. The resulting data model is more general than its project-specific counterparts, and usually represents a least-common denominator approach that only allows some data from each of the contributing projects to be integrated in a useful way. Consequently, data warehouses usually cannot suffice for project-specific data management tasks, because they do not accommodate all of the information contained in any project-specific database.

As an example, consider the VegBank data model. This model allows federation of vegetation plot data for quantifying the composition of plant communities in space and time. The VegBank database contains raw data for vegetation plots, which can be collected in a number of different ways, but is focused on describing the floral composition within an area, along with the environmental context. Plots data form the basis for the classification of vegetation communities, which are associations of co-occurring plant taxa. Data about community classifications are contained within VegBank along with the plots data.

Plots stored in VegBank can have an optional value for the "disturbanceType" to which a plot may have been exposed. The values for "disturbanceType" must be chosen from a controlled list that includes "Animal, general," "Grazing, domestic stock," "Grazing, native ungulates," "Herbivory, invertebrate," and "Herbivory, vertebrates." Individual project data collection methods may not all map precisely into this

Metadata: information used to document and interpret data

breakdown of disturbance types, because the categories are not mutually exclusive and allow for differences in interpretation. Consequently, the data in VegBank may contain less detailed information than the original data sources as a byproduct of being transformed into a more standardized and broadly accessible form. This is a common and unavoidable trade-off for data warehouses. Nevertheless, there are great benefits from integrating the plot data into a common data model because it allows efficient searching for relevant records across a much larger collection and the ability to analyze the data in a common format.

Another limitation of data warehouses is that contributing project-specific data to one warehouse does not automatically make the information available in others. As an example, part of the VegBank plot observation includes a characterization of its soil profile. Contributing these soil data to VegBank, however, does not relate them in any way to the Natural Resources Conservation Service Soil Data Mart from the U.S. Department of Agriculture (**http://soildatamart.nrcs.usda.gov**). Thus, the approach of using vertically integrated databases does not address the issue of integrating with data resources outside the scope of its own data model. Data warehouses essentially have the same data integration problems as project-specific databases, but at a higher level.

Because of this mismatch between the information management needs of individual projects and those of vertically integrated databases, there will always be a cost to contributing data to the latter. In addition, despite the highly integrated nature of data warehouses, they still require extensive documentation of the data to permit their reasonable interpretation. For example the term location in one data set might refer to a very proximate observation such as "on a tree" whereas in another data set it might refer to an entire region, such as "Delaware." Additional metadata—precise, structured descriptions of what a variable is referring to—is needed to resolve these types of issues.

Finally, whether a given scientist is willing to bear the cost in time and effort to contribute their own project-specific data to a data warehouse is driven primarily by their expectations of utilizing the warehouse for their own research purposes. As there is no reward system in place that recognizes the contributions made by sharing data, it is difficult to convince researchers to build and contribute to these data systems simply because there will be a benefit to their scientific discipline.

4.2. Metadata-Driven Databases (Data Collections)

Metadata is the contextual information needed to understand and use a set of data (i.e., data about data). The importance of metadata has been emphasized both within ecology (Jones et al. 2001; Michener 2000, 2006; Michener et al. 1997) and in other communities (Attig et al. 2004, Daniel et al. 1998, Dekkers & Weibel 2003, Nair & Jeevan 2004, Theile 1998, Weibel 1995). Detailed human-readable metadata about the context of data collection, the protocols used to collect the data, and the structure and format of the data objects are a necessary prerequisite to the long-term preservation and interpretation of data. Michener et al. (1997) illustrate this point as a decline in information content with increasing time from when the results of the data are

published. They emphasize that particular events, such as retirement, career change, or death of the original investigator, can have a dramatic impact on the availability of metadata and therefore the utility of data.

An alternative, more robust approach to the highly structured, vertically integrated data warehouse is a more loosely structured collection of project-specific data sets accompanied by structured metadata about each of the data sets. Advantages of this approach include: (*a*) data represented using different data models can be stored together in a single uniform storage system; (*b*) metadata-based data collection is familiar to scientists because it focuses on the same project-level data model as their typical spreadsheet data management approaches; (*c*) the metadata collected is much more detailed than the metadata used in a typical relational database, thus promoting long-term utility of the data; and, (*d*) the metadata is typically more concise than the raw data and can be used as a proxy when searching for data of interest. Each of the data sets is stored in a manner that is opaque to the data system in that the data themselves cannot be directly queried; rather, the structured metadata describing the data is queried in order to locate data sets of interest. After data sets of interest are located, more detailed information (such as the detailed data model that specifies, e.g., the definitions of the variables) can be extracted from the metadata and used to load, query, and manipulate individual data sets.

Some researchers have pointed out that there are limits to what can be captured in metadata and that any attempt to document all aspects of data will necessarily be incomplete and subject to the biases of the original investigators (Bowker 2000). However, this criticism points to the fact that data often reflect implicit assumptions and subtleties in meaning that will be challenging to capture in any structured framework. The increased emphasis on experimental approaches in ecology has certainly resulted in a need for more extensive documentation of methods and procedures in order to reasonably interpret data from these studies. Nevertheless, with the increasing recognition of the value of existing data, even partial descriptions of data that facilitates its reuse beyond that of its original purpose will be important for synthetic studies.

Although, loosely speaking, metadata refers to any information that provides additional context for interpreting data, in practice the term typically has connotations of structured, well-defined categories for systematically documenting critical aspects of a data set, which would be amenable to storage in a data format rather than natural language. A consistent and rigorous set of definitions for metadata categories is called a metadata content specification and, when broadly adopted or endorsed by some community, becomes a metadata standard. Several metadata content specifications could be used for documenting ecological and biological data. Within ecology, the Ecological Metadata Language (EML), developed through the efforts of ecologists and information managers, has garnered support from a variety of institutions such as the National Center for Ecological Analysis and Synthesis (NCEAS), and the Long-Term Ecological Research Network (LTER; Jones et al. 2001). The National Biological Information Infrastructure (NBII) uses the Biological Data Profile (BDP) from the Federal Geographic Data Committee (FGDC; Frondorf et al. 1999, Parr & Cummings 2005). Other metadata standards are in use and development, including

EML: Ecological Metadata Language

LTER: Long-term Ecological Research Network

BDP: Biological Data Profile

the Geographic Markup Language (GML), ISO (International Standards Organization) Geospatial Metadata (ISO 19115), the Directory Interchange Format (DIF), and several taxonomic standards such as the Taxonomic Concept Schema (TCS) from the Taxonomic Databases Working Group (Goodchild 2003). The library community has many metadata standards, including the Dublin Core Element Set for use in providing basic bibliographic information about digital objects (Dekkers & Weibel 2003, Nair & Jeevan 2004). Each of these metadata specifications addresses metadata at differing levels of detail (granularity) that is relevant to biological data.

GML: Geographic Markup Language

ISO: International Standards Organization

KNB: Knowledge Network for Biocomplexity

The multitude of metadata standards creates problems for interoperability. Data providers wanting to make their data available in various national archives often need to comply with multiple metadata standards, which increases the burden on data providers. Partial mappings between metadata standards can be accomplished, but there is no mechanism for automatically synonymizing, or crosswalking metadata concepts among the multiple standards. Within ecology, the Knowledge Network for Biocomplexity (KNB) is a distributed metadata-driven data repository (see Section 4.3 below), which provides a translation from EML to the BDP, but the reverse mapping is not yet available (partly because the BDP metadata is not as fine-grained as EML in several key areas). Another problem with the diversity of metadata standards is that metadata repositories typically support searching via only a single metadata standard, and an integrated search mechanism that spans multiple content standards is still elusive (but see the discussion on EcoGrid in Section 4.3 below).

The level of detail (granularity) and comprehensiveness are important factors for an effective metadata standard. Although several metadata standards provide for basic descriptions of bibliographic information, only a few attempt to fully describe the structure and content of scientific data sets. For example, EML provides a detailed and machine-readable description of the logical structure of data tables (i.e., the data model) as well as their physical structure (data format). This information can be used to automatically parse and load the data sets into analytical systems and relational databases (such as R, SAS, or PostgreSQL), which can significantly improve efficiency of data handling for larger synthesis studies.

Ultimately the choice of a metadata standard for ecological data should depend on the set of capabilities that the metadata provides in terms of facilitating data discovery, access, and reuse for future research, although the relative importance of these might vary depending on institutional priorities and budgets (e.g., federal agencies, NGOs, university researchers). It is possible that ecological metadata may well become accessible in several standards, especially if technology facilitates the interchange among these via automatic translations.

4.3. Current Data Collections

Several national data collections have been created to promote data sharing and long-term preservation of scientific data that are relevant to ecology and biology (**Table 1**). Many are metadata-driven data collections based on one or a few metadata standards. The KNB Metacat system (Berkley et al. 2001) provides a distributed data archive with search facilities that are customized for the specific needs of an

Table 1 Metadata and data collections in widespread use within ecology

Collection name	Standards supported	Archives data
Knowledge Network for Biocomplexity	EML, BDP, others	yes
NBII Metadata Clearinghouse	BDP	no[a]
NSDI Metadata Clearinghouse	CSDGM	no[a]
GBIF Taxonomic Collections	Darwin Core	no[a]
TOPP	EML	yes
Kruger National Park/SAEON	EML	yes
Open-Source Project for a Network Data Access Protocol (OPeNDAP) / Integrated Ocean Observing System (IOOS)	—	yes
VegBank	US NVC	yes
TreeBase	—	yes
Storage Resource Broker	various	yes
ORNL DAAC	BDP, other	yes
Global Change Master Directory	DIF	no[a]
ESA Data Registry	EML	no[a]
Ecological Archives (data papers)	various	yes
Ocean Biogeographic Information System (OBIS)	Darwin Core[b]	no[a]
Global Population Dynamics Database (GPDD)	various	yes

[a]Although this is a metadata clearinghouse, the data are often archived in other systems.
[b]OBIS uses an extension of Darwin Core, see Grassle (2000) for details.

GBIF: Global Biodiversity Information Facility (Taxonomic Names and Specimen Collections)

organization (**Table 1**). This interlinked set of data archives provides a mechanism for investigators to publicly share data or to share only with a limited set of colleagues. Since its inception in 2000, the data holdings in the KNB have been growing rapidly to the current level of over 12,000 databases described in the system (**Figure 1**). In the museum community, the Global Biodiversity Information Facility (GBIF; Canhos et al. 2004) is providing a centralized portal for searching over 450 museum collections via a federated subset of the data that is held in the collections. Both of these resources and others in **Table 1** enable new types of collaborative studies and meta-analysis that were previously impossible.

Several of the systems in **Table 1** only store metadata entries that describe the data and do not store the actual data, whereas others permit archiving the actual data along with the metadata. Registries, which house only metadata, are helpful in increasing the awareness of data holdings, and promote data sharing while bypassing some difficult cultural issues associated with asking scientists to archive (contribute) their data. For example, the Ecological Society of America (ESA) Data Registry (**http://data.esa.org**) was created to better document the data used in articles published in the ESA's journals. The society plans to promote data sharing and preservation by registering (through metadata) the data used for a publication. In future versions of the ESA registry, the society plans to add a data archive feature and integrate its *Ecological Archives* journal into a single integrated system (Peet 1998). The data registries for the Organization of Biological Field Stations (OBFS),

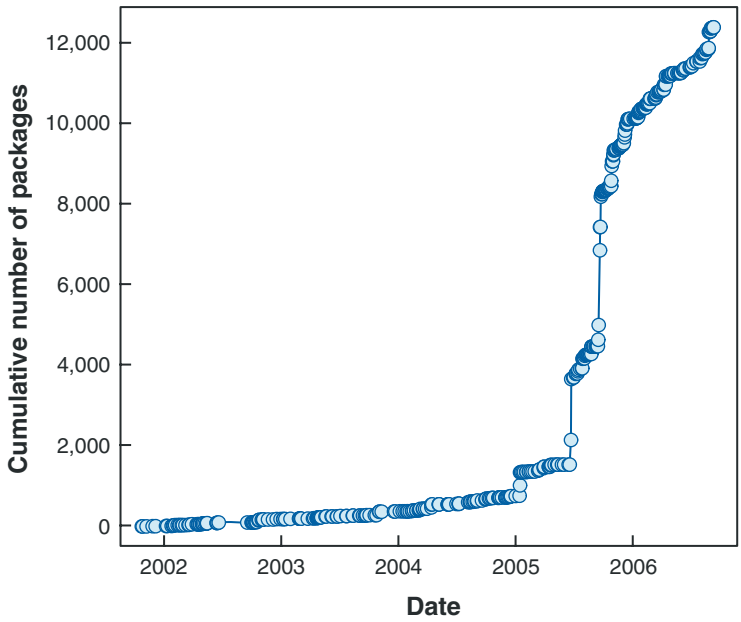

Figure 1

Cumulative number of data packages deposited in the Knowledge Network for Biocomplexity (KNB) over time. Recent advances in data sharing networks such as the KNB have promoted a surge in the number of ecological data sets available. Other systems such as the National Biological Information Infrastructure Metadata Clearinghouse are also growing rapidly (see Parr & Cummings 2005).

U.C. Natural Reserve System (UCNRS), and the NBII are adopting similar strategies of first encouraging scientists to catalog their data and later introducing the idea of archiving the raw data. The OBFS, UCNRS, and ESA data registries are all currently based on the KNB framework.

Data collections must ultimately provide easy access to the actual data, and not just the metadata, if they are to be highly useful in broad-scale synthesis studies. The KNB provides access to data through the use of EML metadata, which describes where to download the data and what format the data will be in. Because the metadata is in a machine-readable format, the process of locating, downloading, and reloading the data can be automated. This automation can significantly increase efficiency for scientists who want to utilize many data sources in a single integrated analysis or model. Other metadata systems use different mechanisms for archiving or associating the data with metadata entries.

Development of a single informatics solution to enable ecologists to gain reliable, long-term data access across these various systems is still a challenge. Recent efforts by the Science Environment for Ecological Knowledge (SEEK) project to create a uniform access interface that works with widely different data systems has seen some success. The SEEK EcoGrid interface currently provides ecologists with

unified access to several of the data collections listed in **Table 1**, including the KNB Metacats, the GBIF taxonomic data portal, and the Storage Resource Broker, as well as to the Geosciences Network (GEON) data portal (Michener et al. 2005; **http://seek.ecoinformatics.org**). This integrated data access system has proved to be extremely useful in the Kepler analytical environment for building analyses and models that utilize heterogeneous data from one or more of these different data systems (Altintas et al. 2004; and see below).

GEON: Geosciences Network

ISBN: International Standard Book Number

ISSN: International Standard Serial Number

LSID: life science identifier

DOI: digital object identifier

4.4. Identification and Versioning of Data

Many of the data collections described in **Table 1** contain entries for the same or overlapping data sets. For example, some of the data found in the KNB are also cataloged in the NBII metadata clearinghouse. Any given data set also likely exists as multiple different versions that represent successive generations of additions, error corrections, or other changes. In addition, integrated data products used in meta-analysis and synthesis studies contain records from multiple primary data sets. Consequently, identifying unique, nonduplicated data objects within and across data repositories is labor intensive.

Standardized data identifiers that are recognized across data collections provide one potential solution for definitively identifying a data set. Standardized identifiers are used by publishers for books (ISBN) and periodicals (ISSN), but this practice is less well established within the electronic data publishing community. Several approaches are emerging within the Internet community for providing unique, location-independent identification of digital data and metadata, including the Life Science Identifier (LSID), Digital Object Identifier (DOI), and other specifications (Clark et al. 2004). These types of identifiers can be used to label unique snapshots of data so that they can be referenced permanently and unambiguously.

The ability to unambiguously reference a specific data set is especially important to the scientific ideal of repeatability. To replicate an analysis, one must reassemble the same data used in the original analysis, which is typically impossible because of current data management practices with respect to error correction and handling of revisions and updates to the data. Relational database systems, which currently provide the main repositories for large amounts of archived biological data, can be complex and dynamic frameworks that are dependent on expert support for continued operation. This can cause problems for future repeatability because these systems may not be available years later when a researcher needs to replicate an analysis. Consequently, good scientific practice requires permanently archiving snapshot versions of a data set in nonproprietary formats with an unambiguous identifier that clearly differentiates that data set from others (see Buneman et al. 2004 for another approach).

4.5. Data Curation

Several technical and nontechnical issues must be dealt with to ensure that data are adequately preserved or curated for future use (National Research Council 1997, Olson & McCord 2000). One of the most pressing issues is establishing a cyberinfrastructure

that supports reliable and long-term data archives (Atkins et al. 2003). Data curation does not come free—it requires appropriate technical infrastructure to which data can be contributed, supporting data providers in the use of this infrastructure, and maintaining the day-to-day operations of the data archives.

The prevalent model for funding of scientific research overlooks the need for long-term preservation of data, with a strong focus on the production of scientific results, and their presentation in the scientific literature. The genomic and molecular biology community is an exception, with well-known and well-supported repositories for the data in the National Center for Biotechnology Information (NCBI). As noted elsewhere, the success of this effort has been facilitated by the relative uniformity and simplicity of sequence data. But the budget to sustain such a data curation center is substantial, amounting to tens of millions of dollars per year. Natural history museums are another notable exception, where digital curation of specimen data is a high priority task (Krishtalka & Humphrey 2000).

Current efforts toward curating data in other areas of biological science are less focused. There is general recognition of the need to preserve the data, but no official or authorized repositories exist for most research data. Despite interest from the digital library community in extending its archiving functions to that of scientific data preservation, scientific data archives are not considered core to their mission (Fox et al. 2002).

The NBII (**http://www.nbii.gov**) provides a registry for metadata that focuses on documenting biological data arising out of federally supported research, as well as from a growing list of agency and academic partners. As part of the United States Geological Survey (USGS), the NBII is subject to the vagaries of budgeting within the USGS, but holds great promise for constituting a robust and persistent archive for biological information. Unfortunately, the current NBII metadata clearinghouse is not a data archive. Moreover, whether the data curated by the NBII contains adequately detailed metadata for future interpretation and scientific reuse needs to be carefully evaluated.

A persistent and reliable data archive, based on well-established standards for metadata and data provisioning, especially for organismal, ecological, and environmental data, remains largely undeveloped. This deficit is also apparent throughout other scientific disciplines, including the physical and social sciences (Atkins et al. 2003, Lord & MacDonald 2003).

The technical needs regarding data curation are similar for all sciences (Freeman et al. 2005, Lord & MacDonald 2003, Raven et al. 1998). The peculiar challenge for the ecological realm is that the field is so broad that diverse types of data are potentially useful for future work (Gross & Pake 1995). Also, although computing and storage capabilities are becoming more affordable, the volume of potentially relevant data is growing even faster. This growth in data volumes is likely to accelerate as new technologies such as dense ground-based sensor networks (e.g., National Ecological Observatory Network; also see Porter et al. 2005) enable exciting new forms of science to be accomplished.

Data curation is not only technically challenging, it can also be expensive. The ongoing costs of data curation break down into three areas: hardware and software,

networking, and staffing. It is difficult to clearly separate costs associated with necessary aspects of cyberinfrastructure (e.g., having a fast Internet connection, or staff to maintain databases and Web servers), from those specifically dedicated to data curation and provisioning. One recent study (Lord & MacDonald 2003) found that staffing was the most significant cost component at several data archiving sites, ranging from 69% to 82% of the total budget.

Maximizing the utility of any investment in data archiving will depend on providing adequate outreach and support so that the data can be effectively discovered and reused by scientists. Finally, clear determination of the cost/benefit of data archiving is difficult owing to the lack of clear metrics for assessing this ratio (Lord & MacDonald 2003). Still, it is undeniable that vast funds are expended on data creation and acquisition. It is false economy, and poor scientific practice, not to ensure that the data are present and useful to all users in the future.

The most cost-effective and accurate way to document data may often involve having the researcher document their own data at the time when they are collected. It will be challenging to find the balance of responsibility for documenting data between individual researchers and trained data stewards who have advanced expertise with appropriate metadata standards and technologies. One hopes that standard approaches to data curation will become common practice once appropriate tools and frameworks are in place (Jones et al. 2001, Michener et al. 1997).

5. DATA INTEGRATION, ANALYSIS, AND MODELING

The previous section described a number of informatics problems—the paucity of metadata and other documentation, incompatibility among different data management systems, the lack of persistent archives, and the need to promote sound data curation practices within the research community. In this section we introduce the possibilities of several emerging informatics technologies for providing major new capabilities to ecological researchers.

5.1. Bridging Data Islands

Much biological information is collected by researchers working relatively autonomously, carefully designing experiments and gathering data that address specific, predetermined hypotheses. This leads to many bioinformatics resources being distributed in data islands, bounded by subdisciplines with their specific focus of interests, specialized vocabularies, and entrenched traditions with regards to informatics (Davis et al. 2005). These data islands present a challenge for recently created synthesis centers in fields such as ecology, evolution, and hydrology, which depend on using existing data for their analyses. There is also a growing realization that more integrative work in biology is not only needed, but increasingly possible owing to emerging informatics solutions that will enable researchers to reach across these islands of data (**http://www.nsf.gov/od/lpa/forum/colwell/rc010324aibs.htm**) to achieve exciting new synthetic results.

Ultimately the rationale for preserving data lies in their potential reuse for addressing new hypotheses, and this usually entails significant data integration challenges.

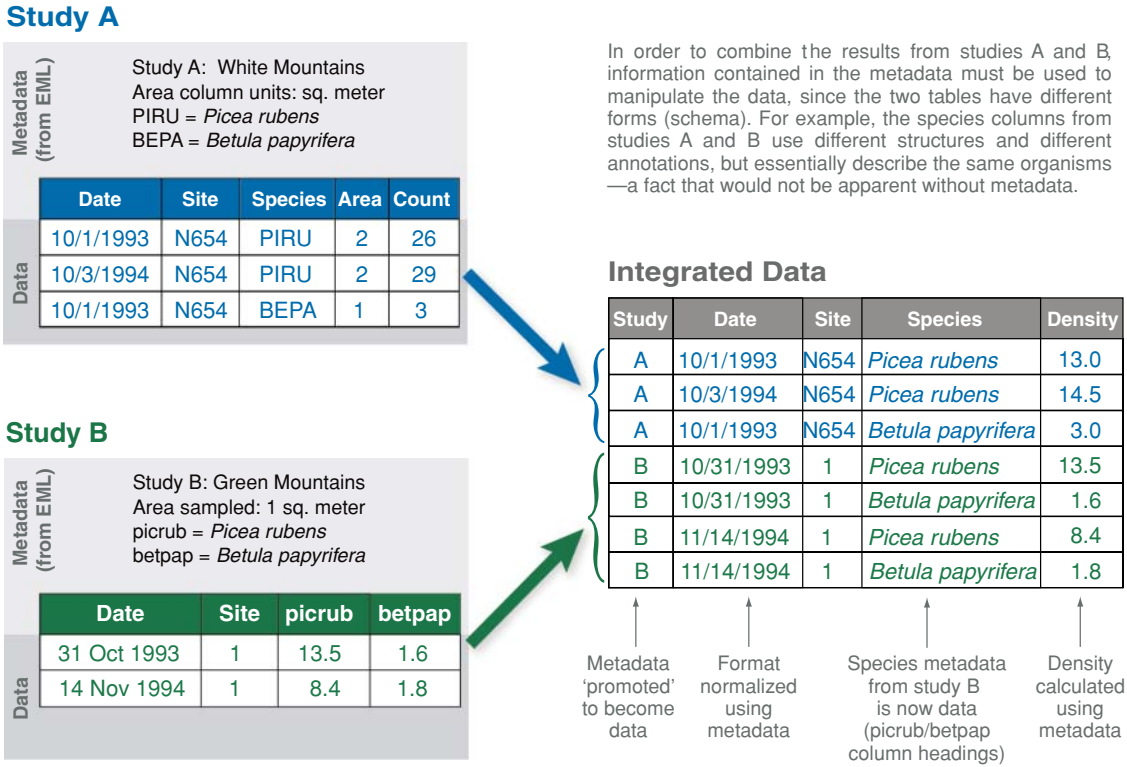

Figure 2

Data integration involves combining two or more heterogeneous but compatible data sets into a uniform product that resolves differences among the source data. Integration requires metadata about each source data set that can provide bridging information, but more importantly requires an understanding of the underlying semantics of the data in order to make reasonable decisions regarding correspondences among the source data sets.

Figure 2 shows two hypothetical source data tables (*left*) that a researcher might want to integrate for use in an analysis (*right*). Performing this integration requires resolving differences in the data format, logical data model, and semantic meaning of data. Although this can be accomplished manually through painstaking expert evaluation of the data sources, such approaches are not practical when investigators need to integrate tens or hundreds of data sources. Thus, automating data integration as much as possible is critical to advances in broad-scale synthesis studies.

Semantics: in computer science, refers to making computers capable of interacting powerfully and appropriately using familiar and meaningful concepts for humans

5.2. Traditional Data Integration Approaches

Data integration involves determining whether and how two or more data resources can be effectively combined. The issue has been widely studied in computer science, resulting in many approaches and systems for managing data differences at the system, format, data model, and semantic levels (e.g., see Haas et al. 2002, Hammer &

McLeod 1999, Ludäscher et al. 2006). Systems-level integration involves reconciling differences in network protocols (e.g., HTTP versus FTP for file transfer), operating systems, and data management applications (e.g., Oracle, Excel, and R). Systems-level integration is necessary for providing low-level support for accessing and transferring data (e.g., data transfers between Windows and Unix), but does not guarantee that an integrated data product is useful or interpretable from a scientific perspective. Format-level integration is similar, but deals with differences in data representation schemes, such as whether the data is stored in a relational (i.e., tabular) or hierarchical database system, or clarifying whether the data object is a raster file or table. Detailed metadata can often enable software applications to cope with formatting issues.

Data-model integration begins with information on how data sources are logically structured. We use the term data model (also called a schema) to refer to a number of essential and concrete features of a data set, e.g., in the case of tabular data, the definitions and data types (integer, string) of the variables (columns) of the table and how those columns might match up with columns from other tables. Structural integration focuses on matching up simple features of data sets, such as like-named variables and checking for consistent data types (e.g., integer or character string). Such information is often explicitly stored within a relational DBMS, but can be lacking in less structured data such as spreadsheets.

One traditional approach to data integration that requires working with data models is data federation (see Haas et al. 2002). In this approach, data integration requires defining a single, global data model (or global schema). Once clarified, the schemas of the local data sources are mapped to the global schema so that one can issue a query against the global schema. The system then uses those mappings to retrieve the associated data from the local databases. This process is a formal description of what one does when creating the vertically integrated database warehouses described above.

The utility of the federation approach is hampered by the difficulty of defining a useful global schema (Batini et al. 1992). Moreover, the federation process cannot be easily extended, because every additional data set to be integrated may require significant modification to the global schema, or lead to major compromises in developing the local schema. An additional potential impediment to federating data via a global schema is that biological databases are often structured for specific usages and have closer affinities with on-line analytical processing (OLAP) and multidimensional databases than a typical relational database (Gray et al. 1997, Pedersen & Jensen 2001, Shoshani 2003). Automated matching of data models (Leser & Naumann 2005, Rahm & Bernstein 2001) as well as peer-to-peer-based integration methods (Bernstein et al. 2002) are also under development to assist in data integration. All of these approaches are likely to benefit from incorporating semantics (Bowers et al. 2004a, Ludäscher et al. 2003, Paton et al. 1999).

5.3. Semantic Approaches to Data Integration

One of the most challenging hurdles to integrating data is uncertainty about its precise meaning—for the data set as a whole, its individual variables (in the case of a table),

Ontology: a formal model of knowledge in a particular subject area useful in making inferences about data

OWL: Web Ontology Language

or even the broader context that motivated its collection. This occurs because the semantics or meaning of some aspects of the data often are unclear. For example, column labels of *wt*, *bm*, and *LL* in separate tables might all refer to a measure of "biomass of leaf litter," but this biologically meaningful concept is not explicit in the abbreviated labels and might not be in any of the metadata. The areas within computer science that deal with clarifying these semantic issues include conceptual data modeling, knowledge representation and semantic mediation, and the Semantic Web (Antoniou & van Harmelen 2004, Berners-Lee et al. 2001, Ludäscher et al. 2003).

Semantic integration involves clarifying data content in ways that are similar to controlled vocabularies, but using more powerful formal structures known as ontologies. Ontologies establish a set of well-defined concepts or terms of interest within a domain and clearly specify how these terms are interrelated (Baker et al. 1999, Brilhante 2003, Horridge et al. 2004, Noy & Hafner 2000, Rector et al. 2004). Owing to the formal logical structure of ontologies, computer-based reasoning systems can use them to draw inferences or conclusions (Gali et al. 2004, Horridge et al. 2004, Sowa 2000). This enables ontologies to help identify important aspects of a data set that were hidden, or implicit. For example, a data set might contain information for a count of organisms and an area over which this measurement was taken. An ontology could be used to identify that density (count divided by area) is implicit in this data set. Other examples of the use of ontologies for integrating ecological data can be found in Bowers et al. (2004b).

The genomics community has made significant advances using ontologies to facilitate data integration by unifying terminologies among communities of genetic researchers working with different model systems (originally fruit fly, mouse, and yeast). These groups of researchers recognized that their work had much in common, but that communication was hindered owing to subdisciplinary variations in terminology. The Gene Ontology project arose as a collaborative effort to develop a structured, controlled vocabulary for associating gene products with their cellular location, molecular function, and biological process, regardless of the model system. The resulting Gene Ontology facilitates a better understanding of the structure and function of genes across taxa (Ashburner et al. 2000).

The creation of robust and useful ontologies requires understanding of formal logic (Baader et al. 2003, Rector et al. 2004), as well as proficiency with specialized software tools (e.g., Protégé; Horridge et al. 2004). The general approach today involves using the OWL Web Ontology Language (McGuinness & Van Harmelen 2004) along with Description Logic reasoners (e.g., Pellet or Racer; Sirin et al. 2005). In ecology, the SEEK (**http://seek.ecoinformatics.org**) and SPIRE (**http://spire.umbc.edu**) projects are developing ontologies to investigate how semantic approaches can assist with data integration (Bowers et al. 2004a,b). These are challenging tasks involving knowledge engineers and computer scientists working closely with biologists to identify and explicate the concepts and relationships that are meaningful for interpreting their data. Although constructing robust ontologies requires considerable skill (Guarino & Welty 2002, Pinto & Martins 2004), and these capabilities are still nascent within the ecological community, we believe that semantic approaches to

data integration are likely to remain an active frontier in bioinformatics for the near future and could be extremely important for a diverse, complicated discipline like ecology.

Scientific workflow: formal, executable model of a quantitative process linking discrete analytical modules; used to document analytical processing, data flow and provenance

5.4. Analysis and Modeling Using Scientific Workflow Systems

As more data become available and techniques improve for integrating datasets, the bottleneck in synthesis increasingly becomes analysis and modeling support. Using spreadsheets and scripted analysis frameworks (e.g., R, SAS, and Perl), scientists are often limited to accomplishing their analyses using the techniques and tools available within the particular software package. For many analyses, however, various techniques and tools are desired across multiple software packages, which invariably results in the process of exporting data from one package and importing it into another. Combining software packages in this way is often challenging and makes tracing and reconstructing the flow of data among applications difficult. As a consequence, information about the analytical procedures that were applied to data is frequently lost. Thus, the process used to obtain a particular analytical result can be difficult to replicate.

Several projects are working to solve these issues by explicitly modeling the flow of data through an entire analytical process (Ludäscher et al. 2005, Osterweil et al. 2006). These scientific workflow systems typically support multiple analytical frameworks and components and have been successfully used in a variety of disciplines, including ecology, the geosciences, molecular biology, and other areas where data access, modeling, and visualization are complex and multistaged (Altintas et al. 2004, Deelman et al. 2004, Ellison et al. 2006, McPhillips & Bowers 2005, Oinn et al. 2000, Pennington & Michener 2005, Taylor et al. 2003).

There are several advantages of scientific workflow systems. First, they provide a formal description of the analytical steps used in a process. Second, they often provide direct access to data sources that would otherwise require significant effort to gather and collate (Shankar et al. 2005). For example, the Kepler scientific workflow system (**Figure 3**; **http://www.kepler-project.org**) is being developed by several scientific communities (Altintas et al. 2004). For ecologists, Kepler provides direct access to ecological data from hundreds of field stations, collections data from various natural history museums, and molecular biology data through services such as GenBank. Third, scientific workflow systems provide a number of tools for managing data including query, discovery, and integration support (Berkley et al. 2005). Fourth, scientific workflow systems typically provide high-level graphical user interfaces for constructing complex analytical processes (see **Figure 3**) and can display analytical results in a clear and intuitive way to scientists.

Workflows are similar to scripted systems like R in that they allow a formal process to be completely specified, but they have the added advantage of being readily understandable to nonexperts. Scientific workflow systems also focus on reusability, e.g., analytical tools from multiple software packages can be incorporated into a single workflow, and workflows can typically be separated into reusable modules,

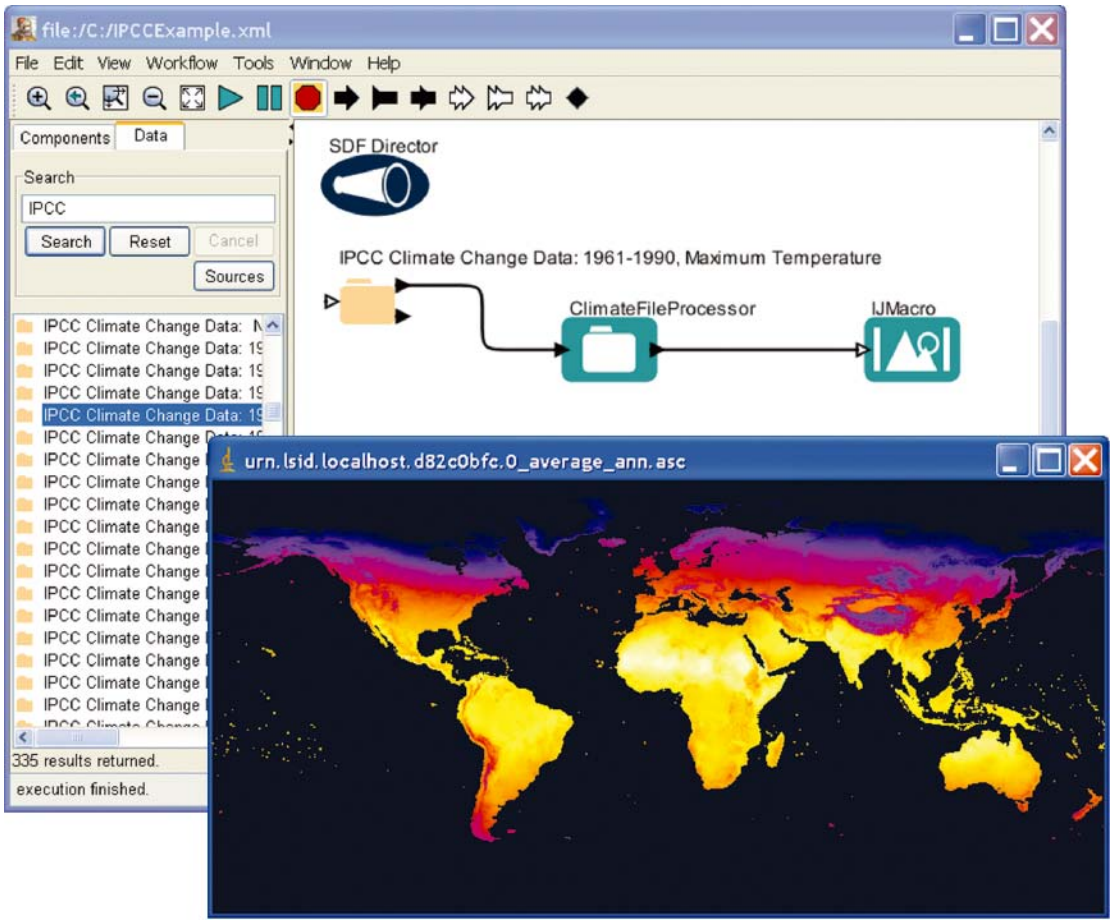

Figure 3

The Kepler scientific workflow system provides a visual model of the flow of data and access to diverse data sources. This workflow shows data from the Intergovernmental Panel on Climate Change (IPCC) that was retrieved from the EcoGrid, processed through a workflow, and visualized as a map.

making complex workflows simpler to understand. Finally, scientific workflows are themselves a form of metadata that can be easily archived and shared with colleagues.

6. CULTURAL ISSUES

There are also nontechnical, cultural, and sociological hurdles to making ecological data broadly available (Palmer et al. 2005). One problem is the reluctance to incur additional operational overhead by learning unfamiliar tools to manage data. This is difficult to overcome even for one's own data, but may be especially onerous when directed toward enhancing the ability of others to use one's data. Many scientists

are reluctant to make their data freely available for fear that others will use it before they themselves have extracted as much as possible for their own analyses and publications. This reluctance is understandable in some cases—imagine conducting a 10-year project during which one makes the data publicly available as it is gathered. It is possible that someone outside the project might use the data available for the first seven or eight years to scoop the primary investigators. However, the opportunities that arise from new collaborations due to data sharing will likely outnumber any unethical use of data, which should be quickly self-correcting owing to the offenders being ostracized and cut out of the flow of funding and scholarship.

Perhaps the most pervasive cultural factor stalling access to digital data is our reward system. Publishing in respectable journals, securing extramural funding, and training the next generation of scholars to do the same are the primary criteria by which we gain the respect of peers and are formally evaluated. Ecological researchers lack an incentive to invest in making their data more broadly available to other scientists because it is neither rewarded nor appreciated and reduces time for respected activities (Olson & McCord 2000). Although it will ultimately benefit science to make data openly accessible, the personal advantages to doing so are not clear in the short term.

Despite these circumstances, Cech et al. (2003) list five principles and 10 recommendations that revolve around the obligation of authors to make data and other materials publicly available. The report stresses that funding entities should require that data gathered under their auspices be made available, that the funding entity provide the resources to do so, and that scientists have a responsibility to the science community to make all data, algorithms, and other information associated with a publication publicly available.

These are admirable goals that have been met, to some extent, in a few disciplines. For example, authors publishing gene sequences must submit their data to GenBank and publish an accession number. GenBank is, however, a relatively simple database of uniform entries compared to the heterogeneity of ecological data. Furthermore, the requirement to publish accession numbers was not an altruistic action on the part of authors or publishers. Rather, publishers could not invest the time and money required to proof and print the massive gene sequences in journals. Currently, there is no such driver to force ecologists to do the same.

Parr & Cummings (2005) comment on how data sharing has transformed other fields and address the points raised above about what has inhibited ecologists from fully participating in the data-sharing revolution. They suggest that it is short-sighted to restrict access to one's data and that the technical barriers to sharing information are illusory. Although not as sanguine about these issues, we believe there are solutions to the cultural hurdles to data sharing.

Ecologists must begin to appreciate efforts made to make data openly available and reward these efforts accordingly. Peers, department heads, deans, and others who evaluate and fund research must recognize that making data openly available is an integral component of research, similar in value to the highly regarded service performed by reviewing journals and grant proposals. We must place a similar value on data sharing.

7. A NEW WORLD OF ONLINE DATA

Major technology advances in ecology have included vast increases in the computing power available to analyze and model patterns and processes, enhanced means to measure and record events and observations, and new methods for exchanging information one-to-one or over the Web. It is ironic then, that the vast majority of ecological data still remain virtually undetectable and inaccessible.

As this situation improves through technological and cultural advances, we can expect changes similar to those occurring with on-line journal publications. For example, citations of articles available electronically are two to five times higher than for those available only in print (T.C. Bergstrom, personal communication). Furthermore, Tenopir et al. (2003) show that the number of articles read and the time spent reading by scientists significantly increases as journals become available electronically. We expect that the changes will be even more dramatic for electronic access to data, because paper journals have been widely available for centuries whereas most ecological data are not accessible in any format. We can also be certain that entirely new ways of using data will emerge, some imaginable (e.g., data analysis might replace reading abstracts or even scientific articles as a means to test ideas for a dissertation or research project) and others beyond our current imagination.

As ecological research becomes more complicated with the addition of critical information from adjacent disciplines (Palmer et al. 2005), access to an ever broader array of data will be indispensable. Although such a capability will advance knowledge about natural systems, it is absolutely essential for the wise conservation and management of natural resources, to the extent that it is unconscionable not to move rapidly toward open access to data. Consider again the headlines of recent articles in *Science* (Kaiser 2003) and *Nature* (Whitfield 2003) about the demise of gorilla populations—"Ebola, Hunting Push Ape Populations to the Brink" and "Ape Populations Decimated by Hunting and Ebola Virus" Although the underlying issue is ecological in nature, the information needed to address this concern comes from many domains, including economics, local culture, disease, and complex population dynamics. It will only be through fundamental advances in informatics that researchers will be able to efficiently access and analyze the diversity of data needed to address such questions in a holistic and comprehensive way. With the clear advantages of online, open access to data, to both our discipline and our planet, we cannot delay in the development and adoption of advanced bioinformatics solutions for enhancing digital access to ecological information.

SUMMARY POINTS

1. There is a critical need for more synthetic and integrative analyses in ecology to better understand and wisely manage the Earth's biological resources.
2. Synthetic analyses in ecology require information from multiple disciplines and perspectives, which raises significant challenges in accessing and integrating relevant data.

3. Although a number of ecological data archives exist, much ecological data is still unavailable.
4. Advances in ecoinformatics are addressing issues regarding data access and integration by adopting semantic approaches and building scientific workflow systems to assist researchers in locating and documenting their data and analyses.

ACKNOWLEDGMENTS

We thank Eric Seabloom, Jim Regetz, and Josh Madin for valuable comments on the manuscript, Leslie Allfree for her administrative support, and Matthew Brooke for assistance with the graphics. Jones, Schildhauer and Reichman are at the National Center for Ecological Analysis and Synthesis, a Center funded by the National Science Foundation (DEB-0072909 and DBI-0225676), the University of California, and the Santa Barbara campus. Bowers is supported by NSF Grant DBI-0533368.

LITERATURE CITED

Altintas I, Berkley C, Jaeger E, Jones M, Ludäscher B, Mock S. 2004. Kepler: an extensible system for design and execution of scientific workflows. *Proc. 16th Int. Conf. Sci. Stat. Database Manag., Santorini Island, Greece*

Andelman SJ, Bowles CM, Willig MR, Waide RB. 2004. Understanding environmental complexity through a distributed knowledge network. *BioScience* 54(3):240–46

Antoniou G, van Harmelen F. 2004. *A Semantic Web Primer*. Cambridge, MA: MIT Press. 272 pp.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. 2000. Gene ontology: Tool for the unification of biology. *Nat. Genet.* 25:25–29

Atkins DE, Droegemeier KK, Feldman SI, Garcia-Molina H, Klein ML, et al. 2003. Revolutionizing science and engineering through cyberinfrastructure. *Rep. Natl. Sci. Found. Blue-Ribbon Advis. Panel Cyberinfrastructure* http://www.communitytechnology.org/nsf_ci_report

Provides a blueprint of the information technology needs for U.S. scientists and outlines the importance of access to data archives.

Attig J, Copeland A, Pelikan M. 2004. Context and meaning: The challenges of metadata for a digital image library within the university. *Coll. Res. Libr.* 65(3):251–61

Baader F, Calvanese D, McGuinness D, Nardi D, Patel-Schneider P. 2003. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge, UK: Cambridge Univ. Press. 574 pp.

Baker P, Goble C, Bechhofer S, Paton N, Stevens R, Brass A. 1999. An ontology for bioinformatics applications. *Bioinformatics* 15(6):510–20

Batini C, Ceri S, Navathe SB. 1992. *Conceptual Database Design: An Entity-Relationship Approach*. Redwood City, CA: Benjamin Cummings

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. 2005. GenBank. *Nucleic Acids Res.* 33:D34–38

Describes the Kepler scientific workflow system and how it uses ontologies to provide advanced data discovery and analysis.

Berkley C, Bowers S, Jones MB, Ludäscher B, Schildhauer M, Tao J. 2005. Incorporating semantics in scientific workflow authoring. *Proc. 17th Int. Conf. Sci. Stat. Database Manag.*, IEEE Comput. Soc.

Berkley C, Jones MB, Bojilova J, Higgins D. 2001. Metacat: a schema-independent XML database system. *Proc. 13th Int. Conf. Sci. Stat. Database Manag.*, IEEE Comput. Soc.

General introduction to the approaches and expected advantages of using ontologies to enhance Web resources.

Berners-Lee T, Hendler J, Lassila O. 2001. The Semantic Web. *Sci. Am.* 284(5):34–43

Bernstein PA, Guinchiglia F, Kementsietsidis A, Mylopoulos J, Serafini L, Zaihrayeu I. 2002. Data management for peer-to-peer computing: a vision. *Proc. Int. Workshop Web Databases (WebDB)*, pp. 89–94

Bowers S, Lin K, Ludäscher B. 2004a. On integrating scientific resources through semantic registration. *Proc. 16th Int. Conf. Sci. Stat. Database Manag.*, pp. 349–52. IEEE Comput. Soc.

Bowers S, Thau D, Williams R, Ludäscher B. 2004b. Data procurement for enabling scientific workflows: on exploring interant parasitism. *Proc. 2nd Int. Workshop Semantic Web Databases (SWDB), LNCS*

Bowker GC. 2000. Biodiversity datadiversity. *Soc. Stud. Sci.* 30(5):643–83

Brilhante VB. 2003. *Ontology and reuse in model synthesis*. PhD thesis. Univ. Edinburgh

Brown JH, Munger JC. 1985. Experimental manipulation of a desert rodent community: Food addition and species removal. *Ecology* 66(5):1545–63

Brunt JW. 2000. Data management principles, implementation, and administration. See Michener & Brunt 2000, 2:25–47

Buneman P, Khanna S, Tajima K, Tan W. 2004. Archiving scientific data. *ACM Trans. Database Syst. (TODS)* 29(1):2–42

Articulates the need for access to more data for effective global modeling of species distributions.

Canhos VP, Souza S, Giovanni R, Canhos DAL. 2004. Global biodiversity informatics: setting the scene for a "new world" of ecological modeling. *Biodiver. Inf.* 1:1–13

Presents a thorough overview of the cultural and ethical issues surrounding data sharing in the life sciences.

Cech TR, Eddy SR, Eisenberg D, Hersey K, Holtzman SH, et al. 2003. Sharing publication-related data and materials: responsibilities of authorship in the life sciences. *Natl. Res. Counc. Rep. Comm. Responsib. Authorship Biol. Sci.* Washington, DC: Natl. Acad. Press

Clark T, Martin S, Liefeld T. 2004. Globally distributed object identification for biological knowledgebases. *Brief. Bioinform.* 5(1):59–70

Connell JH. 1961. The influence of interspecific competition and other factors on the distribution of the barnacle *Chthamalus stellatus*. *Ecology* 42(4):710–23

Daniel R, Lagoze C, Payette SD. 1998. A metadata architecture for digital libraries. *Proc. IEEE Int. Forum Res. Technol. Adv. Digital Libr.*, pp. 276–88. IEEE Comput. Soc., Los Alamitos, CA

Davis MA, Pergl J, Truscott AM, Kollmann J, Bakker JP, et al. 2005. Vegetation change: a reunifying concept in plant ecology. *Perspect. Plant Ecol. Evol. Syst.* 7:69–76

Deelman E, Blythe J, Gil Y, Kesselman C, Mehta G, et al. 2004. Pegasus: Mapping scientific workflows onto the grid. *Proc. Eur. Grids Conf., 2nd, Nicosia, Cyprus*

Dekkers M, Weibel S. 2003. State of the Dublin Core Metadata initiative. *D-Lib Mag.*, 9 (April) No. 4

Ellison AM, Osterweil LJ, Hadley JL, Wise A, Boose E, et al. 2006. Analytic webs support the synthesis of ecological datasets. *Ecology* 87:1345–58

Fox EA, Moore RW, Larsen RL, Myaeng SH, Kim SH. 2002. Toward a global digital library: generalizing US-Korea collaboration on digital libraries. *D-Lib Mag.*, 8(Oct.) No. 10

Freeman PA, Crawford DL, Kim S, Muñoz JL. 2005. Cyberinfrastructure for Science and engineering: promises and challenges. *Proc. IEEE* 93(3):682–91

Frondorf A, Jones MB, Stitt S. 1999. Linking the FGDC geospatial metadata content standard to the biological/ecological sciences. *Proc. 3rd IEEE Comput. Soc. Metadata Conf., Bethesda, MD*, April 6–7

Gali A, Chen C, Claypool K, Uceda-Sosa R. 2004. From ontology to relational databases. *Int. Workshop Concept.-Model Driven Web Inf. Integr. Min., Shanghai, China*

Goodchild MF. 2003. Geographic information science and systems for enviromental management. *Annu. Rev. Environ. Resourc.* 28:493–519

Grassle FJ. 2000. The Ocean Biogeographic Information System (OBIS): an on-line, worldwide atlas for accessing, modeling and mapping marine biological data in a multidimensional geographic context. *Oceanography* 13:5–7

Gray J, Chaudhuri S, Bosworth A, Layman A, Reichart D, et al. 1997. DataCube: A relational aggregation operator generalizing group-by, cross-tab, and subtotals. *Data Min. Knowl. Discov.* 1:29–53

Green JL, Hastings A, Arzberger P, Ayala FJ, Cottingham KL, et al. 2005. Complexity in ecology and conservation: mathematical, statistical, and computational challenges. *BioScience* 55(6):501–10

Good summary of the current possibilities and future needs for more effectively using technology in ecology and conservation.

Gross KL, Pake CE, eds. 1995. *The Future of Long-term Ecological Data. A Report to the Ecological Society of America.* Vol. I: *Text of the Report.* 123 pp. Vol. II: *Directories to Sources of Long-term Ecological Data.* 114 pp. **http://intranet.lternet.edu/archives/documents/other/fledvol2.pdf**

Guarino N, Welty C. 2002. Evaluating ontological decisions with ONTOCLEAN. *Commun. ACM* 45(2):61–65

Haas LM, Lin ET, Roth MT. 2002. Data integration through database federation. *IBM Syst. J.* 41(4):578–96

Hammer J, McLeod D. 1999. Resolution of representational diversity in multidatabase systems. In *Management of Heterogeneous and Autonomous Database Systems*, ed. A Elmagarmid, M Rusinkiewicz, A Sheth, 4:91–117. San Francisco: Morgan Kaufmann. 413 pp.

Horridge M, Knublauch H, Rector A, Stevens R, Wroe C. 2004. *Practical Guide to Building OWL Ontologies Using the Protégé-OWL Plugin and CO-ODE Tools Edition 1.0.* Manchester, UK: Univ. Manchester. 118 pp. **http://www.co-ode.org/resources/tutorials/ProtegeOWLTutorial.pdf**

Jennings M, Faber-Langendoen D, Peet R, Loucks O, Glenn-Lewin D, et al. 2004. Guidelines for describing associations and alliances of the U.S. National Vegetation Classification. Version 4.0. Washington, DC: Ecol. Soc. Am.

Jones MB, Berkley C, Bojilova J, Schildhauer M. 2001. Managing scientific metadata. *IEEE Internet Comput.* 5(5):59–68

Kaiser J. 2003. Conservation biology - Ebola, hunting push ape populations to the brink. *Science* 300:232

Knapp AK, Smith MD, Collins SL, Zambatis N, Peel M, et al. 2004. Generality in ecology: testing North American grassland rules in South African savannas. *Front. Ecol. Environ.* 2(9):483–91

Krishtalka L, Humphrey PS. 2000. Can natural history museums capture the future? *BioScience* 50:611–17

Leser U, Naumann F. 2005. (Almost) Hands-off information integration for the life sciences. *Conf. Innovative Database Syst. Res.*, pp. 131–43

Lord P, MacDonald A. 2003. Data curation for e-science in the UK: an audit to establish requirements for future curation and provision. *JISC Comm. Support Res. (JCSR).* **http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf**

Lubchenco J, Real LA. 1991. Manipulative experiments as tests of ecological theory. See Real & Brown 1991, pp. 715–33

Ludäscher B, Altintas I, Berkley C, Higgins D, Jaeger-Frank E, et al. 2005. Scientific workflow management and the Kepler system. *Concurr. Comput.: Pract. Exp.* Special issue Scientific Workflow

Ludäscher B, Gupta A, Martone ME. 2003. A model-based mediator system for scientific data management. In *Bioinformatics: Managing Scientific Data*, ed. Z Lacroix, T Critchlow, 12:335–70. San Francisco: Morgan Kaufman

Ludäscher B, Lin K, Bowers S, Jaeger-Frank E, Brodaric B, Baru C. 2006. Managing scientific data: from data integration to scientific workflows. *GSA Special Paper 397, Geoinformatics: From Data to Knowledge, ed. A. Krishna Sinha, pp. 109–30.* Boulder, CO: Geol. Soc. Am.

McCullough BD, Wilson B. 1999. On the accuracy of statistical procedures in Microsoft Excel 97. *Comput. Stat. Data Anal.* 31:27–37

McGuinness DL, Van Harmelen F. 2004. OWL Web ontology language overview. *W3C Recommendation*, Feb. 10. **http://www.w3.org/TR/owl-features**

McPhillips TM, Bowers S. 2005. An approach for pipelining nested collections in scientific workflows. *SIGMOD Rec.* 34:12–17

Michener WH. 2006. Meta-information concepts for ecological data management. *Ecol. Inform.* 1:3–7

Michener WK. 2000. Metadata. See Michener & Brunt 2000, pp. 5:92–116

Michener WK, Brunt JW, eds. 2000. *Ecological Data: Design, Management and Processing*. Oxford: Blackwell. 180 pp.

Michener WK, Brunt JW, Helly JJ, Kirchner TB, Stafford SG. 1997. Non-geospatial metadata for the ecological sciences. *Ecol. Appl.* 7:330–42

The seminal paper articulating the need to better preserve ecological data through metadata.

Michener WK, Reichman OJ, Beach J, Jones MB, Ludäscher B, et al. 2005. Creating and providing data management services for the biological and ecological sciences: Science environment for ecological knowledge. *Proc. 17th Int. Conf. Sci. Stat. Database Manag.*, IEEE Comput. Soc.

Morell V. 1996. TreeBASE: The roots of phylogeny. *Science* 273:569

Nair SS, Jeevan VKJ. 2004. A brief overview of metadata formats. *DESIDOC Bull. Inf. Technol.* 24(4):3–11

Natl. Res. Counc., Comm. Issues Transborder Flow. 1997. *Bits of Power: Issues in Global Access to Scientific Data*. Washington, DC: Natl. Acad. Press. 235 pp.

Noy N, Hafner C. 2000. Ontological foundations for experimental science knowledge bases. *Appl. Artif. Intell.* 14(6):565–618

Oinn T, Greenwood M, Addis M, Alpdemir MN, Ferris J, et al. 2005. Taverna: Lessons in creating a workflow environment for the life sciences. *Concurr. Comput.: Pract. Exp.* 18:1067–1100

Olson RJ, McCord RA. 2000. Archiving ecological data and information. See Michener & Brunt 2000, 6:117–41

Osterweil LJ, Wise A, Clarke LA, Ellison AM, Hadley JL, et al. 2006. Process technology to facilitate the conduct of science. *Lect. Notes Comput. Sci.* 3840:403–15

Paine RT. 1966. Food web complexity and species diversity. *Am. Nat.* 100:65–75

Palmer M, Bernhardt ES, Chornesky EA, Collins SL, Dobson AP, et al. 2005. Ecological science and sustainability for the 21st century. *Front. Ecol. Environ.* 3:4–11

Parr CS, Cummings MP. 2005. Data sharing in ecology and evolution. *Trends Ecol. Evol.* 20(7):362–63

Pascal F. 2000. *Practical Issues in Database Management: A Reference for the Thinking Practitioner*. Boston: Addison-Wesley. 256 pp.

One of many good introductions to the conceptual issues underlying data modeling and management.

Paton NW, Stevens R, Baker P, Goble CA, Bechhofer S, Brass A. 1999. Query processing in the TAMBIS bioinformatics source integration system. *Proc. 11th Int. Conf. Sci. Stat. Database Manag.*, pp. 138–47

Pedersen TB, Jensen CS. 2001. Multidimensional database technology. *IEEE Comput.* 34(12):40–46

Peet RK. 1998. ESA journals: Evolution and revolution. *Bull. Ecol. Soc. Am.* 79:177–81

Pennington DD, Michener WK. 2005. The EcoGrid and the Kepler Workflow System: A new platform for conducting ecological analyses. *Bull. Ecol. Soc. Am.* 86(3):169–76

Pinto HS, Martins JP. 2004. Ontologies: how can they be built? *Knowl. Inf. Syst.* 6:441–64

Porter J, Arzberger P, Braun HW, Bryant P, Gage S, et al. 2005. Wireless sensor networks for ecology. *BioScience* 55:561–72

Porter JH. 2000. Scientific databases. See Michener & Brunt 2000, 3:48–69

Rahm E, Bernstein PA. 2001. A survey of approaches to automatic schema matching. *VLDB J.* 10(1):334–50

Raven P, Ayala FJ, Bowker GC, Colwell RR, Cracraft JL, et al. 1998. *Teaming with Life: Investing in Science to Understand and Use America's Living Capital*. President's Comm. Advis. Sci. Technol., Panel Biodivers. Ecosyst., Washington, D.C.

Real LA, Brown JH, eds. 1991. *Foundations of Ecology: Classic Papers with Commentaries*. Chicago: Univ. Chicago Press. 920 pp.

Rector A, Drummond N, Horridge M, Roger J, Knublauch H, et al. 2004. OWL Pizzas: Practical experience of teaching OWL-DL: Common errors and common patterns. *Int. 14th Conf. Knowl. Eng. Knowl. Manag. (EKAW)*. Northamptonshire, UK: Whittlebury Hall

Shankar S, Kini A, DeWitt DJ, Naughton J. 2005. Integrating databases and workflow systems. *ACM SIGMOD Rec.* 34(3):5–11

Shoshani A. 2003. Multidimensionality in statistical, OLAP, and scientific databases. In *Multidimensional Databases: Problems and Solutions*, ed. M Rafanelli, pp. 46–68. Hershey, PA: Idea Group

Sirin E, Parsia B, Grau BC, Kalyanpur A, Katz Y. 2005. *Pellet: a practical OWL-DL reasoner.* **http://www.mindswap.org/papers/PelletJWS.pdf**

Sowa JF. 2000. *Knowledge Representation: Logical, Philosophical, Conceptual Foundations*. Pacific Grove, CA: Brooks Cole. 594 pp.

Taylor I, Shields M, Wang I, Rana O. 2003. Triana applications within grid computing and peer to peer environments. *J. Grid Comput.* 1:199–217

Tenopir C, King DW, Boyce P, Grayson M, Zhanf Y, Ebuen M. 2003. Patterns of journal use by scientists through three evolutionary phases. *D-Lib Mag.* 9(May) No. 5

Theile H. 1998. The Dublin Core and Warwick Framework. *D-Lib Mag.*, Jan. 4. **http://www.dlib.org**

Thomas CE, Ganji G. 2006. Integration of genomic and metabonomic data in systems biology: are we 'there' yet? *Curr. Opin. Drug Discov. Dev.* 9(1):92–100

Weibel S. 1995. Metadata: the foundations of resource description. *D-Lib Mag.*, July 1. **http://www.dlib.org**

Whitfield J. 2003. Ape populations decimated by hunting and Ebola virus. *Nature* 422:551