DEPARTMENTS

Emerging Technologies

Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation.

Introduction

The long-term value and the utility of ecological data for advancing ecological understanding and solving important environmental problems depend on the availability of suitable and adequate metadata, or descriptive information describing data content, context, quality, structure, and accessibility (Michener et al. 1997). As a discipline, ecology is moving beyond its tradition of small-scale empirical observations and experiments conducted by one or a few investigators at relatively small scales (Palmer et al. 2004, 2005). The need to expand the temporal and spatial scales of ecological research necessitates increased data sharing and mechanisms to enable long-term community access to data (e.g., Olson and McCord 2000, Andelman et al. 2004) and presents new challenges for integration of heterogeneous ecological information across a range of spatial, temporal, and organizational scales (Andelman and Willig 2004).

Historically, investigations of many ecological phenomena, and the development of theory to explain them, have been limited by the availability of suitable long-term data. For example, collecting adequate wildlife population data suitable for population dynamic research is extremely time and resource intensive. As a result many population dynamics research activities have focused their collection and analysis on individual datasets. This in turn makes it difficult to formulate general theory and to investigate large-scale spatial and taxonomic patterns. In response to this limitation, and motivated by the need to provide comprehensive access to biological population data, the Center for Population Biology at Imperial College, Silwood Park, the National Center for Ecological Analysis and Synthesis (NCEAS), and the University of Tennessee collaborated to develop the Global Population Dynamics Database (GPDD; ‹http://cpbnts1.bio.ic.ac.uk/gpdd/›). The GPDD facilitates the discovery of general patterns and principles, advances the understanding of large-scale spatial and temporal patterns, and enables researchers to acquire large numbers of datasets, without having to undertake repetitive, time-consuming, and expensive searches. The GPDD is now the largest collection of animal and plant population data in the world, and brings together nearly five thousand time series in one database. It provides an important resource for ecologists, resource managers, and environmental scientists interested in the dynamics of natural populations, or in asking comparative questions about the nature of population variability (e.g., Kendall et al. 1998, Kendall et al. 2000, Fagan et al. 2001, Inchausti and Halley 2001, 2002, 2003).

Synthetic efforts such as development of the GPDD are extremely valuable, but very time consuming and difficult, due to the heterogeneity and dispersion of ecological data, as well as the general lack of adequate data documentation. Ecological data exhibit a range of formats, reflecting different underlying motivations for data collection, different suites of

variables, and different spatial and temporal sampling designs. In addition, ecological metadata typically vary in extent, detail, and quality (Regan et al. 2002, Andelman et al. 2004), and may consist of mental notes, hand-written notes in a field notebook, a comments field in an Excel spreadsheet, or other types of informal documentation. Currently, there exist few standards to guide decisions about what quantity and quality of metadata are sufficient to enable data that initially were collected for a single, relatively narrow purpose, to be understood and used appropriately for a variety of purposes. Unfortunately, this often means that the value of ecological data diminishes over time, because important details about the data are forgotten or lost by the original investigator, because of career changes, or changes in data storage and management technology (Michener 1997).

Here we describe Ecological Metadata Language (EML; ‹http://knb.ecoinformatics.org/software/eml/›), a method for formalizing and standardizing the set of concepts that are essential for describing ecological data. We explain why and how documenting metadata with EML will extend the long-term utility of ecological data and facilitate the processes of data discovery and integration.

Metadata

Metadata is the information that describes "who, what, where, when, why, and how" an ecological dataset was collected. Metadata is simply data about data. Most ecologists have experienced the difficulty in remembering important details about their own data, even after only a few months have passed since it was collected. Unless data are adequately documented, this difficulty only increases over time. Even the simplest analysis requires some level of metadata. For example, consider a simple data table, with no column headers (Table 1). Without metadata, a data table such as this one is useless. Unless we know the definitions of the columns and their measurement units, the numbers are meaningless. Furthermore, in this example, there is no metadata to identify the location where data were collected, the focal organism or system, or the identity and location of the data owner.

Table 1. Ecological data with no metadata.

VO	5/30/2002	1	*AVFAT*	4.25	3.19	0.01
VO	5/30/2002	1	*BRHOR*	5.33	3.19	0.01
VO	5/30/2002	1	*CALUT*	3.33	3.19	0.01

Table 2 illustrates the same data table, with some additional metadata. In this example we can see that data were collected in May 2002 at a site identified as VO. However, one can only guess at what VO might signify, or at the meaning of the columns with headers "Sp," "Bm," "P," and "N." Additionally, the values of the columns ("P" and "N") are measurements that are taken at the *plot level*. As a result these values are repeated for each species observation in a plot. This may potentially cause problems if an ecologist incorrectly assumes each record contains an independent observation for these two data columns. Thus Table 2 contains plot level and species level (within-plot) data mixed together in a given observation. This situation is quite typical in ecological datasets, which are frequently formatted more for analysis than efficient storage (such as in a relational model). But in general, the lack of metadata makes this dataset relatively unusable by anyone other than the original owner.

Table 3 describes the same dataset as in Table 2, but with more comprehensive documentation. The data owner is identified, column headers are defined, and general information is provided regarding how and where the data were collected. Access to this information greatly clarifies many aspects of this dataset. However, the original data owner probably has additional information that could further enhance the utility of these data for a broader range of research activities in the future. For example, inclusion of geographic coordinates or other information that describes the spatial location where the data were collected, or information about codes or numbers used to indicate missing data, would increase the potential utility of these data.

This example illustrates another common feature of metadata: the information provided is "whatever"

Table 2. Ecological data with a limited amount of metadata.

Site	Date	Plot	Sp	Bm	P	N
VO	5/30/2002	1	*AVFAT*	4.25	3.19	0.01
VO	5/30/2002	1	*BRHOR*	5.33	3.19	0.01
VO	5/30/2002	1	*CALUT*	3.33	3.19	0.01
VO	5/30/2002	2	*AVFAT*	20.82	11.91	0
VO	5/30/2002	2	*BRHOR*	30.22	11.91	0
VO	5/30/2002	2	*CALUT*	25.62	11.91	0
VO	5/30/2002	3	*AVFAT*	6.00	9999	9999
VO	5/30/2002	3	*BRHOR*	7.11	9999	9999
VO	5/30/2002	3	*CALUT*	9.11	9999	9999
VO	5/30/2002	4	*AVFAT*	4.56	5.14	0.22
VO	5/30/2002	4	*BRHOR*	12.36	5.14	0.22
VO	5/30/2002	4	*CALUT*	11.34	5.14	0.22
VO	5/30/2002	5	*AVFAT*	6.17	7.15	0.35
VO	5/30/2002	5	*BRHOR*	5.68	7.15	0.35
VO	5/30/2002	5	*CALUT*	7.16	7.15	0.35
VO	5/30/2002	6	*AVFAT*	4.80	14.35	0.15
VO	5/30/2002	6	*BRHOR*	6.70	14.35	0.15
VO	5/30/2002	6	*CALUT*	9.06	14.35	0.15
VO	5/30/2002	7	*AVFAT*	23.44	12.65	0.45
VO	5/30/2002	7	*BRHOR*	36.55	12.65	0.45
VO	5/30/2002	7	*CALUT*	17.04	12.65	0.45
VO	5/30/2002	8	*AVFAT*	3.45	4.42	0.76
VO	5/30/2002	8	*BRHOR*	4.11	4.42	0.76
VO	5/30/2002	8	*CALUT*	6.24	4.42	0.76
VO	5/30/2002	9	*AVFAT*	2258	3.55	0.76
VO	5/30/2002	9	*BRHOR*	19.58	3.55	0.76
VO	5/30/2002	9	*CALUT*	27.878	3.55	0.76
VO	5/30/2002	10	*AVFAT*	14.56	18.53	0.91
VO	5/30/2002	10	*BRHOR*	17.45	18.53	0.91
VO	5/30/2002	10	CALUT	19.56	18.53	0.91

the data owner decided to document. Without standards or guidelines for metadata content, if the same individual were to document another dataset, the same or different information might be included, and the format might or might not be the same. In this way, if analyses require multiple datasets from different owners, locations or time periods, it is unlikely that all relevant datasets would have metadata with equivalent levels of detail, use consistent terminology, or use consistent formats for metadata. This suggests the need to standardize metadata. Table 4 provides an example of a much more detailed metadata document that was made using EML, in which each metadata concept (from dataset title to geographic description) has been formalized and standardized. Note that this table is merely one possible "view" of EML metadata.

Creating metadata with EML

Ecological Metadata Language (EML) is a method for formalizing and standardizing the set of concepts that are essential for describing ecological data. EML grew from an open-source, community-based effort involving ecological researchers, information managers, and software developers, led by the National Center for Ecological Analysis and Synthesis (NCEAS) and the Long Term Ecological Research Network (LTER). The need for EML or a similar method to facilitate preservation and long-term utility of the growing archives of ecological data has been recognized for some time (FLED Report 1995, Michener et al. 1997, Olson and McCord 2000). Other metadata standards such as the Federal Geographic Data Committee Biological Data Profile (FGDC BDP) exist and are either currently interchangeable with EML or will be in future development efforts. EML is currently more comprehensive in some aspects than the BDP (e.g., the data entity and attribute metadata fields).

Table 3. Relatively comprehensive, but unstructured metadata.

Written metadata

This experiment was designed to collect productivity, diversity, and soil data for Northern California grasslands. The results were published in a paper titled, "Soil nutrients and the relationship between diversity and productivity" (Doe and Smith 2003). Data were collected at two sites, the Coastal Hills Reserve and the Valley Oak Reserve, within the coastal mountains of Northern California. The area is primarily oak (*Quercus* spp.) savannah and grasslands on limestone soil. In spring of 2002, 10 1-m^2 plots were randomly distributed throughout a 100-km^2 area of each location. All plots were placed on grasslands. In each plot, plants were identified to species and then clipped at root level, dried, and weighed to obtain aboveground peak standing biomass. As most of the production is from annual plants, peak standing biomass can be used as an approximate measure of annual productivity. Approximately 0.5 g of soil was collected from the midpoint of each plot. This soil was taken back to the laboratory and analyzed for total nitrogen and phosphorus content.

All three species were observed in the plots. Nonnative plants observed included: *Avena fatua* and *Bromus hordeaceus*. Native plants included *Calochortus lutens*.

Codes used in data table (Table 2) are given below:

Site: Site at which data were collected – VO = Valley Oaks Reserve; Date: Date data were collected, mm/dd/yy format; Plot: Randomly assigned number of plot; Sp: Species code for each species found in plots.

Species name	Code
Avena fatua	AVFAT
Bromus hordeaceus	BRHOR
Calochortus lutens	CALUT

Bm: Biomass, measured in grams for each species; P: Phosphorus in soil, recorded in ppm (parts per million) per plot; N: Nitrogen in soil, recorded as a percentage per plot.

Data were collected by PI Jane Doe with assistance from graduate student John Smith in conjunction with the staff of the Coastal Hills Reserve and the Valley Oak Reserve. Collection of the data was funded by NSF grant No. 12345. Data may be used freely. Please acknowledge persons, grants, and reserves in any resulting publications.

Contact information:
Jane Doe
Department of Biology
Northern California University
University Town, CA 95666
(321) 654-0987
E-mail: doe@uncal.edu

EML is intended for use by any ecologist or manager of ecological information. It describes a range of essential aspects of ecological data, such as the names and definitions of variables; units of measurement; date, time and location of data collection; the identity of the individual who collected the data; sampling design; etc. EML attempts to reduce ambiguity and uncertainty by formalizing these metadata concepts into a comprehensive yet standardized set of terms and definitions intended specifically for ecological data. The metadata in Table 4 provides an example of a dataset that has been reasonably well documented using EML.

The question of "How much metadata is enough?" does not have a clear-cut answer. There are two factors to consider: the effort involved in creating the metadata, and the value derived from it afterwards when trying to discover or interpret the data. In general, assume that "more is better," because omitting detail from metadata at the outset may lead to problems later on (e.g., hours of discussion or exploratory analyses), and in the worst case may render the data unusable. We have found that once researchers are familiar with the basics of ecological metadata, they can create EML for an overall dataset (ownership, contact information, motivation, spatiotemporal context, key words, etc.) in about 30 minutes. Provision of detailed descriptions of the variables (attributes) and their definitions and units can be more time consuming, depending on the number of variables and the possibility of complicated interactions of datasets among one another. However, the more metadata an ecologist creates, the longer and more usable their dataset will be for future research. Indeed, the utility of the metadata may actually increase through time, as advanced tools emerge for automatically processing datasets based on their metadata. For now, once a scientist understands some basic aspects of metadata, they can derive a decent understanding of a moderately complicated dataset after reviewing the metadata for about 20 minutes.

A walk through the EML metadata shown in Table 4 will clarify some of the more important metadata concepts provided in EML. The information in Table 4 is arranged in five broad metadata sections, each of which contains more detailed metadata. These sections are intended to categorize EML metadata fields in a way that is intuitive to ecologists. The metadata categories include the General Dataset, Geographic, Temporal, Taxonomic, Methods, and Data Table Metadata sections.

The General Dataset metadata section contains concepts that identify and name the dataset and describe the purpose of the data collection and the questions the data were originally intended to address. These "discovery" fields allow for searching for the data in various computer-based data catalogs. Some types of metadata, such as the title and abstract, are self-explanatory, but others may not be. The *usage rights* field provides a place for information about who can ethically and legally use the dataset, and what, if any, restrictions there are on usage. Other general dataset metadata information includes contact information for people who had a significant role in collecting or managing the data, such as the dataset *creator* and a dataset *contact*. The *contact* should be the person to whom further questions regarding the data and metadata should be addressed. Funding information, such as a grant number or acknowledgement of a private donor, should also be documented here. Additionally, EML provides fields for entering bibliographic information pertaining to analyses based on the data. EML supports a range of standard reference styles.

As the name suggests, the Geographic Metadata section is used for geographic and spatial metadata. The geographic description field contains information about where the research project took place, where samples were collected, and any spatial or geographic references that may provide a context for the data. Latitude and longitude may also be described to increase geographic accuracy. This field is optional in EML, as some ecological datasets may not have a strong geographical context, such as laboratory-generated data or model output.

The Temporal Metadata section contains information about when the data were collected. Information

can be stored either as a range of dates (e.g., data were collected every month between June 2002 and 2003) or specific time periods (e.g., 1 May 2002, 08:00–12:00 and 1 June 2003, 08:00–12:00). In addition, information about potential gaps in data collection or in the collection of some variables can be included.

If the dataset has species information, the Taxonomic Metadata section can be used to describe the species. Information such as the taxonomic authority (i.e., the book or system that is used to identify a species) and the taxonomic rank (i.e., Family, Genus, Species) can be described here.

Sometimes individual data collection efforts may be part of a larger research project. For example, a large research project involves data collection at a number of field sites where the larger research projects goals and objectives, as well as personnel (e.g., principal investigators, data managers, etc.) may vary from those at the individual field sites. EML has a Protocol section that attempts to capture the important metadata at the larger research project level, whereas the EML Methods section documents the implementation at the field site level. For example, a PI at one of the sites decides not to implement the larger research projects standard protocol per se, because of local conditions (e.g., a species is known to occur at much lower densities, so more quadrats are needed than at other sites). Thus, the Protocol section contains the standardized sampling design that all sites are intended to implement, and the Methods section records what *actually* takes place when the data are collected at the field site. The Methods section also describes things such as the machines or devices used to collect data, types of quality control used to ensure data are measured and recorded correctly, and the spatial units of the samples being collected. The Methods section should be sufficiently detailed to allow someone to recreate the data collection efforts. Unlike the Methods section of a published article, there is no need to be terse with respect to metadata—fully detailed descriptions can be quite helpful here.

The Data Table section is useful when the data are in a tabular format (rows and columns). There are fields for physical information such as the file name, whether or not the values in the data table are case sensitive, the number of records, and the structure of the data table (i.e., attribute names in columns or rows). This category also contains metadata regarding the columns of data themselves. The *name* field contains a unique name for the column in the table and is usually very short (it is often the column header). The *label* is a more descriptive word or phrase that describes the column and is useful because acronyms or ambiguous abbreviations often are used as column headers. The *definition* field contains a definitive description of the column, indicating what the values in the column represent and how those values relate to the methods described in the Methods section. The *unit* and *type* fields contain the units (grams, meters) and data types (e.g., integer, floating point, etc.) for each column. *Missin*g documents the number or symbol used to indicate that no data were collected (e.g., 9999). *Precision* is often useful to document, e.g., when the numbers in a column represent output from a machine, there is often some level of precision associated with the data. The *attribute domain description* provides definitions of any codes used in that column (e.g., VO = Valley Oak) and the domain of values that are valid in the column (e.g., biomass values have a *domain* of real numbers greater than zero, while the actual observed range might be from 3.33 g/m^2 to 36.55 g/m^2).

The metadata contained in Table 4 represents a level of detail that would be highly useful to someone with little or no prior knowledge about the dataset (or the dataset owner after not working with the data for a few years) to determine whether or not the data are appropriate for some intended use. Additionally, once someone has decided to use the data for a particular purpose, the metadata should be sufficient to enable the next research steps (e.g., contact the data owner for the dataset, or if the data are public and accessible, begin preliminary analyses). This sort of detailed information about the attributes—their definitions and units—is particularly useful to any analyst hoping to explore patterns in the dataset.

Table 4. Metadata for: diversity, productivity, and soil data for North American grasslands. (This is one possible rendition of the EML contents for this dataset.)

General metadata		
Abstract	This research program was designed to collect productivity, diversity, and soil data in Northern California grasslands. Data collected include species richness, presence/absence of plant species, peak standing biomass, and nitrogen and phosphorus soil content. The relationship between diversity and productivity can take many different shapes. Soil nutrients can affect species composition, diversity, and productivity. This research program will attempt to investigate soil nutrients as a possible factor in determining the shape of the diversity/productivity curve. Funding was used to collect richness, plant presence/absence data, biomass, and soil nutrient data.	
Keywords	Richness, productivity, grasslands, biomass, northern California, soil nutrients.	
Usage rights	This dataset is publicly available through the Knowledge Network for Biocomplexity. ‹www.knb.ecoinformatics.org› Please acknowledge the Knowledge Network for Biocomplexity, NSF Grant No.12345 and Dr. Jane Doe in any publications that use this data.	
Funding	NSF Grant No.12345	
Individual: Owner	Dr. Jane Doe	
Position	Associate Professor	
Address	Department of Biology, Northern California University, University Town, CA 95666 USA	
Phone	(321) 654-0987 (voice)	
E-mail	doe@uncal.edu	
Web address	‹http://www.uncal.edu/doe›	
Individual: Primary contact	John Smith	
Position	Data manager	
Address	Department of Biology, Northern California University, University Town, CA 95666 USA	
Phone	(321) 654-0987 (voice)	
E-mail address	smith@uncal.edu	
Web address	‹http://www.uncal.edu/smith›	
Organization	Department of Biology, Northern California University	
Article citation		
Author	J. Doe	
Author	J. Smith	
Date	2002	
Title	Diversity, productivity, and soil nutrients at a Northern California grassland	
Journal	Plantae	
Volume	23	
Issue	2	
Page range	1–10	
Geographic metadata		
Geographic description	Data were collected in the coastal mountains of Northern California, in the Valley Oak Reserve. The Valley Oak Reserve is adjacent to and managed by Northern Caifornia University (NCU). NCU is located in University Town in Sonoma County, ~150 km northeast of San Francisco. The Valley Oak Reserve is located on the east-facing slope of the California coastal mountains. The area is primarily oak (*Quercus* spp.) savannah and grasslands. The reserve is 100 km² in area	
Bounding coordinates	Longitude	-120º 5’00: degrees -120º 5’00: degrees
	Latitude	39º00’00: degrees 38º45’00: degrees

Temporal metadata		
Temporal description	The observations in the data were made in the late spring, at approximately peak biomass. Observations were made during the following range of dates:	
Begin	05/30/2002	
End	05/30/2002	
Taxonomic metadata		
Taxonomic authority	Type	Book
	Author	W. L. Jepson
	Editor	J. C. Hickman
	Date	1993
	Title	The Jepson manual: higher plants of California
	Publisher	University of California Press
	Publication place	Berkeley
General information	All herbaceous plant species were recorded.	
Classification	Taxonomic level	Family
	Taxonomic name	Poaceae
	Taxonomic level	Genus, species
	Taxonomic name	*Avena fatua*
Classification	Taxonomic level	Family
	Taxonomic name	Poaceae
	Taxonomic level	Genus, species
	Taxonomic name	*Bromus hordeaceus*
Classification	Taxonomic level	Family
	Taxonomic name	Liliaceae
	Taxonomic level	Genus, species
	Taxonomic name	*Calochortus lutens*

Methods metadata

General sampling design

Ten 1-m² plots were randomly placed throughout the Valley Oak Reserve. Due to destructive biomass harvest, plots are relocated each year.

Two plastic sample bags (Ziplock) are labeled with the randomly assigned plot number and contents (plant or soil). If more than one bag is needed, all bags are labeled with the contents, plot number and bag number (i.e., the second of four bags of plant clippings from plot 6 will be labeled "Plant, Plot 6, Bag 2/4")

Species diversity and biomass measurement

At each plot, one person, starting at the southeast corner of the plot, identifies each plant according to the Jepson manual. Species names are recorded in the field notebook.

Plant material for each species within each 1-m^2 plot is clipped at soil level and placed in the sample bag.

Sample bags containing plant matter are brought to the laboratory. If wet, plant matter is dried using paper towels.

Plant material is dried in a drying oven at 80°C for 24 (±2) hours.

Plant matter is weighed within 2 hours of drying.

Soil data measurement

Approximately 0.5 g of soil, free from plant debris, is collected from the middle of the plot.

Soil is placed in appropriate sample bag.

Soil samples are placed into aluminum sample trays and placed into a drying oven and dried at 80°C for 24 (±2) hours.

Soil is ground until powdery using a ball mill, and weighed.

Soil sample is analyzed for phosphorus and nitrogen using a SoilPro v. 10 machine. All procedures for this machine are followed.

Quality control

All sampling was done by Jane Doe and John Smith. Jane Doe trained John Smith in how to identify species and calculate biomass and soil data measurements. Several "calibration" plots were used prior to the experiment to ensure methods were appropriate and data collectors each followed them accordingly.

Data table metadata

File name	EML_simple_example.txt
Case sensitive?	No
Number of records	30
Orientation	The data are arranged with major variables in columns.

Data table structure and attribute description

Attribute name	Label	Definition	Unit	Type	Missing	Precision	Attribute description
Site	Site	Site at which data were collected		Integer			Enumerated Code: VO Def.: Valley Oak Reserve
Date	Date	Date data were collected	mm/dd/yy	Date			
Plot	Plot	Randomly assigned 1-m² plot numbers		Integer			Numeric Min.: 1.0 Max.: 10.0
Sp.	Species	Species codes		Text			Enumerated Code: AVFAT, BRHOR, CALUT Def.: *Avena fatua*, *Bromus hordeaceus*, *Calochortus lutens*
Bm	Biomass	Peak standing biomass in plot per species	g/m^2	float			Numeric Min.: 3.33 Max.: 36.55
P	Phosphorous		ppm	float	9999	±0.05	Numeric Min.: 3.19 Max.: 18.53
N	Nitrogen		Proportion (of total soil mass)	float	9999	±0.01	Numeric Min.: 0.0 Max.: 0.91

How to document data with ecological metadata language (EML)

Many tools can be used to create valid EML documents because EML documents are just text documents in a standardized format. Here we highlight two mechanisms targeted at ecologists for creating EML to document ecological data: Morpho and web registries. Future development efforts will eventually allow these mechanisms to create metadata documents in any number of standardized formats (e.g., FGDC BDP). Both of these mechanisms are supported by a very active open source community, which provides user support and helps guide development efforts.

Morpho

Morpho is a data and metadata management software program that works on Windows, Macintosh, and Unix. It enables an ecologist to create, edit and manage metadata and data tables, and is intended to be useful for individual scientists trying to manage their own research data. Consider that today, a researcher typically uses the native file system of their computer to "manage" their data (e.g., nested folders in Windows or Macintosh machines, full of datasets searchable via a potentially cryptic "filename"). This method works in the short term, but becomes problematic as researchers' data holdings increase, and they begin

to forget where they stored something, what it was called, and what it contained. Morpho also provides special capabilities to search and query publicly accessible ecological data archives based on EML (e.g., those in the Knowledge Network for Biocomplexity; ‹http://knb.ecoinformatics.org›). These archives in turn will be tightly integrated with other metadata repositories such as the National Biological Information Infrastructure Metadata Clearinghouse. Morpho includes user-friendly "wizards" that facilitate using a subset of EML (e.g., Table 4) to document the most fundamental aspects of ecological metadata. In addition, Morpho provides access to the entire contents of EML, which currently include over 2000 metadata concepts or terms for describing ecological data. For more information see ‹http://knb.ecoinformatics.org/software/morpho›

Web registries

As interest grows in documenting and preserving ecological data via EML, institutions are starting to create easy ways for their affiliated scientists to accomplish this through the Web. Any ecologist is welcome to document their data using a subset of EML through the web registry at ‹http://knb.ecoinformatics.org/index.jsp› To use this tool an ecologist must first register (i.e., provide basic information about him or herself and how they may be contacted). She can then create basic EML compliant metadata without installing and learning Morpho. As when using Morpho, these EML metadata can become broadly available to the ecological community through the Internet, facilitating the discovery of an ecologist's dataset by other ecologists around the world. Additionally, large organizations and research projects may want to create customized EML web interfaces, as has been done for the Long Term Studies Section of the ESA ‹http://knb.ecoinformatics.org/knb/style/skins/ltss› and the National Center for Ecological Analysis and Synthesis ‹http://data.nceas.ucsb.edu› Currently, the web registries provide a mechanism for creating and querying metadata over the Internet. However, they do not provide direct access to the data unless an online location (such as a URL) is provided with the metadata. Repositories, which archive both the metadata *and* the data, are more effective for long-term preservation of data and access to data, than web registries, which only store metadata and an (optional) link to the original dataset.

The future

By systematically documenting data in a standardized and structured format, ecologists will advance ecological knowledge by contributing to a powerful community-wide data resource that will inform analyses undreamt of at this time. As these ecological data and metadata archives grow, their value will increase. EML provides a *common structure* for these resources, to better enable ecologists to document, share, and interpret ecological data. The formal structure of EML will also facilitate the development of advanced software applications that can process these metadata. EML is implemented in XML (Extensible Markup Language), a growing standard for marking up documents on the Web ‹http://www.w3.org/XML› This means that a wide range of software will become available for manipulating EML metadata, ranging from basic search and query tools that can be used remotely through the Web, to tools for remote integration of heterogeneous datasets, and their analysis and visualization (e.g., ‹http://seek.ecoinformatics.org› More detailed information on EML, tools for creating metadata, and analytical and synthetic development efforts utilizing EML are described at ‹http://knb.ecoinformatics.org/index.jsp›

Acknowledgments

This work and many of the tools mentioned in the manuscript are supported by the National Center for Ecological Analysis and Synthesis (National Science Foundation, Grant No. DEB-007290), the Knowledge Network for Biocomplexity (National Science Foundation, Grant No. DEB99-80154) and the TEAM Initiative of Conservation International, made possible by the Gordon and Betty Moore Foundation.

Literature Cited

Andelman, S. J., C. M. Bowles, M. R. Willig, and R. B. Waide. 2000. Understanding environmental complexity through a distributed knowledge net-

work. BioScience **54**:240–246.

Andelman, S. J., and M. R. Willig. 2004. Networks by design: a revolution in ecology. Science **305**: 1564–1565.

Fagan, W. F., E. Meir, J. Prendergast, A. Folarin, and P. Karieva. 2001. Characterizing population vulnerability for 758 species. Ecology Letters **4**: 132–138.

Gross, K., C. E. Pake, and Members of the FLED Committee. 1995. Report of the Committee on the Future of Longterm Ecological Data (FLED), Ecological Society of America. Ecological Society of America, Washington, D.C., USA.

Inchausti, P., and J. Halley. 2001. Investigating long-term ecological variability using the global population dynamics database. Science **293**:655–657.

Inchausti, P., and J. Halley. 2002. The long-term temporal variability and spectral colour of animal populations. Evolutionary Ecology Research **4**: 1033–1048.

Inchausti, P., and J. Halley. 2003. On the relation between temporal variability and persistence time in animal populations. Journal of Animal Ecology **72**: 899–908.

Kendall, B. E., O. N. Bjornstad, J. Bascompte, T. H. Keitt, and W. F. Fagan. 2000. Dispersal, environmental correlation, and spatial synchrony in population dynamics. American Naturalist **155**: 628–636.

Kendall, B. E., C. Briggs, W. W. Murdoch, P. Turchin, S. P. Ellner; E. McCauley, R. M. Nisbet, and S. Wood. 1998. Why do populations cycle?: a synthesis of statistical and mechanistic modeling approaches. Ecology **80**:1789–1805.

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

Olson, R. J., and R. A. McCord. 2000. Archiving ecological data and information. Pages 117–141 *in* W. Michener and J. Brunt, editors. Ecological data—design, management and processing. Blackwell Science, Malden, Massachusetts, USA.

Palmer, M., et al. 2004. Ecology: ecology for a crowded planet. Science **304**:1251–1252.

Palmer, M. A., et al. 2005. Ecological science and sustainability for the 21st century. Frontiers in Ecology and the Environment **3**:4–11.

Regan, H. M., M. Colyvan, and M. A. Burgman. 2002. A taxonomy and treatment of uncertainty for ecology and conservation biology. Ecological Applications **12**:618–628.

Eric H. Fegraus, Sandy Andelman, Matthew B. Jones, and Mark Schildhauer
National Center for Ecological Analysis and Synthesis
735 State Street, Suite 300
Santa Barbara, CA 93101-3351 USA