Completing the data life cycle: using information management in macrosystems ecology research

Janine Rüegg1*, Corinna Gries2, Ben Bond-Lamberty1, Gabriel J Bowen1, Benjamin S Felzer3, Nancy E McIntyre4, Patricia A Soranno5, Kristin L Vanderbilt6, and Kathleen C Weathers9

An important goal of macrosystems ecology (MSE) research is to advance understanding of ecological systems at both fine and broad temporal and spatial scales. Our premise in this paper is that MSE projects require integrated information management at their inception. Such efforts will lead to improved communication and sharing of knowledge among diverse project participants, better science outcomes, and more transparent and accessible (ie “open”) science. We encourage researchers to “complete the data life cycle” by publishing well-documented datasets, thereby facilitating re-use of the data to answer new and different questions from the ones conceived by those involved in the original projects. The practice of documenting and submitting datasets to data repositories that are publicly accessible ensures that research results and data are available to and usable by other researchers, thus fostering open science. However, ecologists are often unfamiliar with the requirements and information management tools for effectively preserving data and receive little institutional or professional incentive to do so. Here, we provide recommendations for achieving these ends and give examples from current MSE projects to demonstrate why information management is critical for ensuring that scientific results can be reproduced and that data can be shared for future use.

In a nutshell:

- Large research collaborations distributed across space, time, and disciplines require careful documentation of the scientific process, from beginning to end
- As scientific research expands the scales of analysis and synthesis, data re-use becomes vitally important; information management is critical when combining data from various sources
- Additional incentives, support, and training are needed to encourage scientists to publish data that are well-documented in terms of their origin, accuracy, and any manipulations

1Division of Biology, Kansas State University, Manhattan, KS *(jrueegg@ksu.edu); 2Center for Limnology, University of Wisconsin-Madison, Madison, WI; 3JGCRI, Pacific Northwest National Lab, College Park, MD; continued on p 30

www.frontiersinecology.org © The Ecological Society of America
Data management in MSE research

The data life cycle includes the description and preservation of data. A traditional project (dark blue background) includes planning, data collection, data quality control, and analysis. Projects relying on existing data (light blue background) for all or part of their analyses go through the steps of planning, collecting, quality assurance and quality control (QA/QC), additional data discovery, data integration, and finally analysis. To complete the data life cycle (white background), one must add the steps of data documentation (metadata) and data archiving in a publicly accessible repository.

Figure 1. The data life cycle includes the description and preservation of data. A traditional project (dark blue background) includes planning, data collection, data quality control, and analysis. Projects relying on existing data (light blue background) for all or part of their analyses go through the steps of planning, collecting, quality assurance and quality control (QA/QC), additional data discovery, data integration, and finally analysis. To complete the data life cycle (white background), one must add the steps of data documentation (metadata) and data archiving in a publicly accessible repository.

Two steps that should be adopted by the scientific community to complete the data life cycle and ensure the long-term availability and re-use of this material are “describe/document” and “preserve/publish” (Figure 1). Data must be associated with metadata that describe the “how, what, when, where, and who” and then archived so that they remain available (Michener et al. 1997; Whitlock 2011). This enables the data to retain value beyond the life of a project, creating additional opportunities for research. Although the concept of a closed data life cycle is not new, data documentation and publication are particularly critical in facilitating research at broad spatial and temporal scales, such as that associated with MSE research, in addition to allowing the steps “Discover”, “Integrate”, and “Analyze” (light blue section in Figure 1) to be integrated across many projects.

Rigorous data management requires resources and funds for each project. The value of the data can increase throughout and beyond the termination of a project but only if data are properly described, preserved, and made available for future research projects. Therefore, there is clearly immense value in the inclusion of financial support from funding agencies and institutions for data management and preservation activities (Kueffer et al. 2011). In addition, researchers should receive credit (from hiring and promotion committees as well as funding agencies) for data publication as an intellectual contribution to the scientific enterprise (Weltzin et al. 2006). Unfortunately, few scientists are satisfied with the current levels of data sharing or long-term archiving, due in part to lack of funding to gain the expertise needed to properly manage data (Tenopir et al. 2011).

Here we provide examples that demonstrate how incorporation of and interaction with professionals in the environmental information management field are essential at every step of a given project’s data life cycle. Specifically, we highlight strategies in four important areas that are being used in current MSE projects: (1) data collection, (2) integration of data from many sources, (3) data integration across scales and from modeling, and (4) provision of access to and documentation about data.

Data collection: incorporating information management early

Integrating information management into a project early on, before and during the data collection phase, is particularly important when a large volume of data is collected and/or the data are complex, when many people in different places are involved in the data collection, and/or when the data are collected over an extended time period. It is imperative that data are collected through methods recognized by the scientific community (WebPanel 1). Quality control and data aggregation
Data management in MSE research

J Rüegg et al.

approaches need to be transparently documented and data should be labeled according to a standardized vocabulary. Without this degree of coordination, it is difficult to integrate data and assess the quality and fitness thereof for use. Two current MSE projects – “Climate Forcing of Wetland Connectivity in the Great Plains: An Exploratory Study Using Graph Theory” (Wetland Connectivity) and “Grassland Sensitivity to Climate Change at Local to Regional Scales: Assessing the Role of Ecosystem Attributes vs Environmental Context” (EDGE) – exemplify the advantages of addressing data management early in a project’s life cycle.

Both projects involve the collection of large amounts of high-frequency sensor data (e.g., air and soil temperature, carbon dioxide flux, land cover) in addition to other parameters across multiple grassland sites in the US (McIntyre et al. 2014). They differ in that the Wetland Connectivity project is using data collected by other entities (including the US Geological Survey and US Fish and Wildlife Service) whereas the EDGE project is placing sensors in the field and managing the raw data, with the additional task of monitoring sensor performance for high-quality data collection. However, both projects require data to be quality controlled, stored, and secured. The Wetland Connectivity project is using previously curated data with “cloud technology” (i.e., online commercial technologies for storing, securing, and sharing the information) and manages the analytical products but does not store the raw data (this is managed by other entities). In contrast, the EDGE project relies on a sensor network and stores and shares data on a local server.

The EDGE project must develop automated approaches for quality control of the high volume of streaming sensor data. EDGE participants are using a Data Toolbox for MATLAB (The MathWorks Inc, Denver, Colorado) developed by the Georgia Coastal Ecosystem LTER site (Sheldon 2008). This system has many built-in functions for performing data manipulation tasks that would otherwise require custom coding, and provides a user-friendly graphical interface for applying quality assurance/quality control (QA/QC) rules. The toolbox generates a log file of all operations performed for inclusion in the metadata to document any data transformations. EDGE also employs the Open Source DataTurbine (OSDT) server—a real-time streaming data engine that receives data from sensors and then transmits that data to other programs, such as Real-Time Data Viewer (RDV) – to examine sensor performance (Fountain et al. 2012). The RDV provides an interface for viewing time-synchronized plots of the data, thus facilitating the detection of anomalous patterns that indicate sensor malfunctions (Figure 2; Daugherty et al. 2011).

The researchers managing these two MSE projects aim to provide preprocessed data to their collaborators and to demonstrate the importance of well-conceived information management strategies. To allow for meaningful analysis by project members, we argue that community-developed – preferably peer-reviewed – and well-documented standards for gap-filling, aggregating, converting, and modeling of data need to be followed. Data should be provided using defined, controlled vocabularies for variable names to ensure a high degree of clarity and eventual automation of analyses. These information management strategies can enable project participants to access and store preprocessed datasets in a central location with well-documented provenance for access and determination of usability for subsequent analyses by collaborators from different disciplines.

Data integration: integrating data from many sources

Complete databases on ecological variables that span broad spatial and/or temporal scales are rare (e.g., Fitter and Fitter 2002; Bond-Lamberty and Thomson 2010). The datasets that would be included in such databases either do not exist or are too often hidden away in the computers or filing cabinets of many different researchers.
Data management in MSE research

The compilation of such datasets into a single, integrated database is one strategy to conduct MSE research across broad scales. By definition, each dataset originates from a different source but is of the same general thematic type (e.g., lake chemistry but not stream biota). In creating an overarching database, it is assumed (though not always true) that the heterogeneous datasets found in ecology share, or can be converted to, common variables, contain spatially explicit location data, and include information about general methods for sample collection and processing.

Broad-scaled, integrated ecological databases compiled from discrete, individual datasets have enormous value. First, the database provides the necessary ecological observations for new analyses by extending the spatial and temporal extents beyond those commonly studied. With sufficient numbers and distribution of comparable datasets, and appropriate analytical approaches (Gurevitch and Hedges 1999), new research questions that span multiple scales can be addressed (a key aspect of MSE research; Heffernan et al. 2014). Additionally, the database can be integrated with biogeochemical, geophysical, and climatic datasets already available at broad scales (Figure 3). Moreover, statistical analyses using such data may uncover subtle patterns or effects only visible because of the large number of observations, the extent of which may be beyond the capability of any one study to collect.

However, common data integration problems that complicate database compilation include the lack of a consistent coordinate reference system, taxonomic naming inconsistencies, and the semantics of variable names.

Developing integrated databases from individual datasets can be greatly facilitated by federated data repositories (e.g., DataONE; Michener et al. 2012), the use of standard metadata and data exchange formats (e.g., WaterML; Zaslavsky et al. 2007), existing database formats (e.g., CUAHSI ODM; Tarboten et al. 2008), the development of workflow tools (Jones and Gries 2011; Michener and Jones 2011), and the development of ontologies (WebPanel 1; Madin et al. 2008). Such information and tools can help ameliorate the problem of integration, but not (yet) automate it. Integration is still extremely time consuming and cannot typically be achieved with off-the-shelf software. The development of such an integrated database is the goal of the MSE project “The Effect of Cross-Scale Interactions on Freshwater Ecosystem State Across Space and Time,” in which a database of lake chemistry will be constructed using data collected from approximately 15,000 lakes across 17 US states. The data were obtained from state, tribal, and federal natural resource agencies, university researchers, citizen groups, and environmental consulting agencies. Each of the steps in creating an integrated database presents its own challenges:

- Discover potential datasets: datasets are often not publicly available and have to be identified based on insider knowledge.
- Obtain datasets: generally, the dataset owner should be identified and contacted, and data access and use, including the form of acknowledgement, must be negotiated.
- Develop a database schema: data integrity and logical consistency must be ensured, but the inclusion of different types of data requires flexibility.
- Develop a strategy for tracking data provenance: the origin and metadata for each dataset need to be integrated into the database, documenting data manipulation steps and QA/QC approaches.
- Integration into one data model and comparable measurement units: importing datasets into the schema requires both programming skills and domain knowledge.

Figure 3. Description of how multiple individual ecological databases distributed across broad geographic areas can be integrated with broad-scaled, single-themed databases to create a multi-themed macrosystems ecology database.
It is critical that informatics professionals and ecologists collaborate during all steps of data consolidation. Data integration across datasets highlights some of the issues that would be addressed through improved documentation and the use of metadata at the individual project level by knowledgeable individuals, such as information managers.

### Data integration: from local observations to models and back

Many MSE research projects need to integrate large-scale data products with site-based observational data. One area with particularly high levels of demand for such products is in large-scale biogeochemical and ecological modeling. Typical input variables for global biogeochemical models (e.g., the Terrestrial Ecosystem Model [Hayes et al. 2011] or the Community Land Model [Oleson et al. 2010]) include locally measured fields such as surface-level ozone (O₃) indices and nitrogen deposition, land use and cover, and climate data products based on decades- to centuries-long monthly or daily point measurement datasets (e.g., Kistler et al. 2001). However, datasets differ in resolution and accuracy and to prepare them all as input to a particular model at a predetermined resolution provides an information management challenge. The MSE project “The Future of Ecosystems and Extremes: Using Diverse Environmental Data Sets in Support of Regional to Global Earth-System Models and Predictions” will be using different datasets as inputs to several biogeochemical models to determine the ecosystem response to climate extremes. A goal of this project is to reduce model uncertainty due to model structure by comparing ecological outputs from different models based on common meteorological and environmental input conditions.

Difficulties involved in developing gridded (i.e., cell-based) datasets from point-based observational and model data include, among others, spatial and temporal interpolation. Interpolation involves scaling spatially from sites to grids, scaling temporally from longer (i.e., monthly) to shorter (i.e., hourly) timescales, and considering how to condense information available at the subgrid scale (Reilly et al. 2012; Levy et al. 2014). For instance, monthly or seasonal indices for O₃ (e.g., SUM06 or AOT40 indices) are often used to represent the detrimental effects of O₃ on vegetation and must be developed from O₃ data collected hourly (Felzer et al. 2004, 2005). Gridded data products (e.g., Climatic Research Unit gridded data products; Mitchell et al. 2004), as well as tools for developing such products, are now available from a diverse and growing array of sources, including modern sensor networks and historical data. New research continues to lead to the development of improved data products, such as the atmospheric and land-use data representing the eddy covariance footprint measured at Ameriflux (http://amerifluxornl.gov) or NEON (www.neoninc.org) sites. Several recent projects have combined eddy covariance data with either remote-sensing coverage from Moderate Resolution Imaging Spectroradiometer or biogeochemical model data to provide gridded datasets of gross primary productivity and net ecosystem exchange for the contiguous US at resolutions of 1 km (Xiao et al. 2008, 2010, 2011) and for the globe at half-degree resolution (Jung et al. 2009, 2011). These types of datasets provide an invaluable resource for modelers to validate their ecosystem function output or optimize model parameters via Bayesian approaches (Tang and Zhuang 2009), thus reducing uncertainty in these parameters. Currently, most gridded data products exist as managed resources that are produced, updated, and distributed by specific research and monitoring groups and agencies (e.g., the National Atmospheric Deposition Program, http://nadp.sws.uiuc.edu). While management by specific groups works well for creating standardized, quality-controlled data products, such products are unique to a dataset and not necessarily flexible in terms of accessibility and use.

Research projects, including MSE studies, that aim to improve model representation of how climate change affects ecosystems will rely heavily on these datasets, and it is particularly important to carefully document procedures used for interpolation and/or aggregation, as well as the provenance of incorporated data. Users of any datasets, gridded or otherwise, must be able to judge a dataset’s fitness for use for the question under consideration and to be able to assess data quality based on this documentation.

### Completing the data life cycle: documentation and sharing facilitate analysis

There are several ongoing efforts to develop cyberinfrastructure that facilitates access to data resources. Some data repositories are providing direct access to data within statistical packages and workflow systems (e.g., see WebPanel 1) via web-based services, facilitating streamlined analysis of data and documentation of procedures, while others are developing more specialized analytical tools that are available online along with the data. For instance, the Isoscape Modeling, Analysis, and Prediction (IsoMAP) toolkit, which focuses on environmental isotope data and is being developed by a collaborative team of ecologists, Earth scientists, information managers, computer scientists, and statisticians, provides grid-supported geospatial analytical capacity linked directly to diverse data collections. The resulting resources are being integrated into the MSE project “Inter-university Training for Continental-scale Ecology” in support of research and graduate education. IsoMAP will be used as a platform, making spatial data analysis and modeling accessible to students in intensive interdisciplinary courses where the diversity of student backgrounds and expertise would prohibit hands-on research with many other toolkits (Bowen et al. 2012b).

Examples of the application and use of IsoMAP vary in scope and level of user knowledge. Among IsoMAP’s most widely useful aspects is a geostatistical toolkit that supports predictive modeling of continuous spatial fields and creation of gridded data products. Applications of
this tool include converting regional observations of groundwater isotope ratios into raster maps showing the contribution of re-evaporated lake water to precipitation (Bowen et al. 2012a) and mapping continental-scale patterns of variation in the hydrogen isotope ratios of precipitation that can be used to discern patterns of bird migration based on measurements of the same isotope in feathers (Hobson et al. 2012).

Development and use of data resources within IsoMAP encompasses many of the information management practices introduced above. The data have been compiled from multiple sources by the development team, and each data subset is documented by metadata stored within a metadata catalogue (Bowen et al. 2012b). IsoMAP implements automated data-processing workflows to facilitate the extraction, manipulation, and preparation of data requested by users, and queries to the metadata catalogue allow this system to identify and retrieve appropriate datasets. The users’ interaction with the data processing system is simplified through a set of interactive, browser-based workflow components (Figure 4). These simplifications increase the accessibility and efficiency of IsoMAP geoprocessing operations but also limit the system’s flexibility relative to desktop geographic information system applications.

The provisioning of data analysis and advanced visualization tools through an interface like IsoMAP improves open science and collaboration. Analyses can be standardized and documented because IsoMAP data resources and tools are versioned and all processes conducted within the system are automatically documented and archived in metadata. Finally, IsoMAP analyses are conducted on NSF-XSEDE (www.xsede.org) grid computing resources. Although current IsoMAP tools exploit a small fraction of the computing power available, the potential exists to drastically increase analytical complexity and data intensity as additional functionality is developed and implemented.

### Conclusions

Macrosystems research involves the collection of large amounts of raw data, mobilization of previously unavailable data, scaling of data products, and custom analytical tools, as well as challenges in data validation, documentation, visualization, and storage. Robust information management must be part of any MSE research project, and the full engagement of specially trained personnel is indispensable for project success. While training researchers in data management is necessary, the inherent complexity of the diverse tools and products necessitates that skilled professionals be included on MSE teams. Publishing project data, as well as aptly rewarding team participants for such publication, should be part of a project’s definition of success and supported by both funding agencies and institutions (Goring et al. 2014). Recognizing the importance of data as the foundation upon which science is built clearly demonstrates that information managers are key members of a scientific team. Data documentation and preservation are critical; scientific data are irreplaceable, particularly in a changing global environment, and will likely become, either alone or as part of a future integrated analysis, the basis for new research and discoveries.

### Acknowledgements

We thank all participants at the NSF-MacroSystems Biology PI meeting in Boulder, CO (March 2012), for fruitful discussions that led to this Special Issue and this paper. In particular, J O’Neil-Dunne provided comments on the inception and outlines of this manuscript. We also thank H Gholz and L Blood (NSF) for helpful discussions, as well as the MacroSystems Biology Program in the Emerging Frontiers Division of the Biological Sciences Directorate at NSF for support. For author contributions, see WebPanel 2.

### References


3Department of Geology and Geophysics and Global Change and Sustainability Center, University of Utah, Salt Lake City, UT; 7Department of Earth and Environmental Sciences, Lehigh University, Bethlehem, PA; 4Department of Biological Sciences, Texas Tech University, Lubbock, TX; 5Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI; 6Department of Biology, University of New Mexico, Albuquerque, NM; 7Cary Institute of Ecosystem Studies, Millbrook, NY

www.frontiersinecology.org © The Ecological Society of America