Integrating Ecological Data: Tools and Techniques

John H. PORTER
University of Virginia, Charlottesville, VA 22903 USA

and

Kenneth W. RAMSEY, Jr.
New Mexico State University, Las Cruces, NM 88003 USA

ABSTRACT

Integration of data is critical to achieving new levels of understanding of ecological systems and processes. Typically, data integration is achieved only through a painstaking manual process that rules out large-scale integration. We believe that many of the techniques related to uncertain reasoning (fuzzy logic, Baysian networks, and evolutionary algorithms) and data mining might be usefully applied to ecological data integration. Here we present two case studies. One characterizes a traditional approach to integration. The second focuses on using software system integration to integrate geospatial and research data, along with providing data discovery services. We discuss those case studies where advanced techniques might prove useful and where modifications are needed to support scientific research.

Keywords: Data Integration, GIS, Uncertain Reasoning, Data Mining, Information Management Systems, Ecoinformatics

INTRODUCTION

Advancement of ecological science is increasingly dependent upon our ability to integrate data from diverse sources. The understanding of ecological processes at the spatial scales of the landscape, region and globe and at the temporal scales of the decade, century and the millennium require data that span these scales. Such data go beyond the collection abilities of any individual investigator or single research project and thus require data integration [1].

Although need to integrate diverse ecological and environmental data is not new, the opportunity to do so is. Traditionally, ecologists have not shared data, nor have they had adequate incentives to do so [2]. Traditionally, data has been collected, analyzed and publications prepared by a single individual or small group of researchers, typically a professor and associated graduate students. As discussed by Strebel et al. [3] and Michener et al. [4], over time most of this data have been lost through a slow process of "data decay" as, in the absence of metadata, our ability to locate or interpret data has been diminished or lost. However, in the last decade there have been sociological and technological developments that have led to increased sharing of data. These include the role of the Internet and World-Wide Web in lowering the costs of sharing data [5], the implementation of information management policies by research projects that define the responsibilities of data providers and data users [2], the recognition by funding organizations such as the National Science Foundation that data products, as well as publications, are valuable results from research projects and even the development of "data journals" such as the Ecological Society of America's new journal "Ecological Archives."

With ecological data becoming more readily available, the critical issue becomes not how we can get it, but rather what we can do with it. In this, ecology is not alone. As was stated in a recent *Science* article: "Computer technology has facilitated the collection of data so well that now, in a growing number of fields, the availability of data is no longer (or soon will not be) the limiting factor for addressing fundamental scientific questions. Paradoxically, the new limitation is computer technology: Only with the help of computer science can we make sense of the masses of data that computers have enabled us to collect, and share and discuss the data with colleagues around the globe. The challenge now is to design aids to help us comprehend data so complex or interconnected that we cannot organize, integrate, or understand it alone" [6]. Here we present two case studies: an example of project-specific data integrationand an example of system-based integration of spatial and thematic data and discuss research directions for developing techniques to address large-scale integration needs.

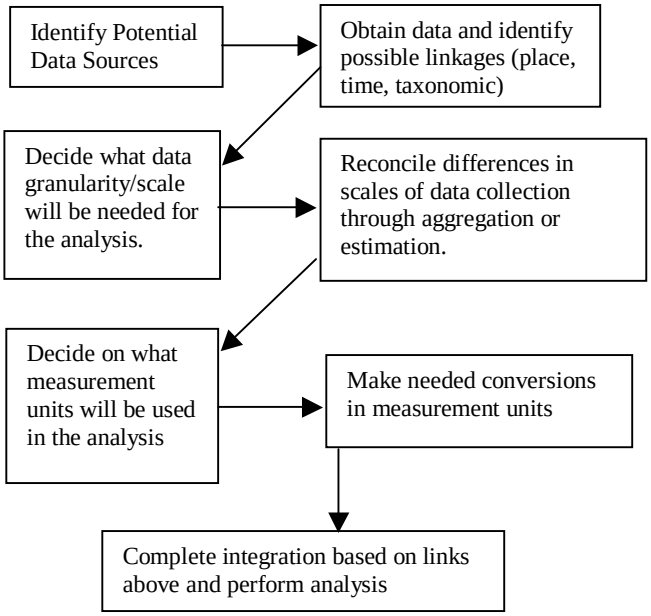
Scientific data systems are required to deal with data that is both more complex and incorporates more (necessary) inconsistencies than traditional business databases [7-9]. However, the development of "data warehouse" and "data mart" systems in business [10], coupled with the development of standards for ecological metadata [4] diminish these differences in important ways and lead to opportunities for cross-fertilization of disciplines.

CASE STUDY: PROJECT-SPECIFIC INTEGRATION

Project-specific integration of ecological data remains

primarily a manual process (Figure 1). Once a scientific hypothesis is defined, an investigator or group of researchers will identify possible datasets, either datasets that they maintain themselves, can obtain from published literature or (more recently) can download from an on-line database. Once obtained, decisions are made about parameters that will be used to link data and about the scale and level of aggregation. For example, if some data are collected on an hourly basis while other data are collected only once per year, the hourly data must be aggregated to provide an annual value that can then be merged 1:1 to the one-per-year data. The process of integration demands painstaking concentration on maintaining data quality as misleading results can occur if errors are introduced at any point in the analytical process.

Figure 1: The Data Integration Process



The identification of linkages is a key step in data integration. For ecological data, the most important linkages tend to be temporal and spatial. Every observation was made at a point in space at a particular time, although the availability and representational form this information takes may be highly variable. Taxonomic linkages are also possible, although evolving taxonomic standards often makes this difficult because species names are not invariant over time [11].

A study underway at University of Virginia investigates the relationship between specific meteorological and climatological factors and primary production of vegetation as part of the Virginia Coast Reserve Long-term Ecological Research (VCR/LTER) project [12]. Here data on climate and productivity is being integrated based on indices of time and space. After discussion, the investigators determined that, although hourly and daily meteorological measurements were available, monthly and annual aggregations of data were the most likely to yield results, as most of the productivity data are at those time scales. The ongoing integrative analysis depends on individual investigators using a suite of traditional software tools, such as statistical packages and spreadsheets, individually preparing temporally-indexed data structures (Table 1). Although software is used, the process remains primarily manual, with each decision being made by experts familiar with a specific portion of the data and the final integration being performed by a group of scientists during the course of intensive analysis sessions.

Table 1: Software used for project-specific integration

Software	**Use**
WWW browsers (Internet Explorer, Netscape)	Data discovery and download
Spreadsheet Software (Excel, Quatro)	Data entry and display, some limited graphing. May also be used for some final analyses using data imported from statistical packages
Statistical Packages (SAS, SPSS)	Data merging, analysis and graphics
Teleconferencing (NetMeeting, Polycom and other H.323 compatible products)	Communication between collaborating researchers. Especially important is T.120 application sharing capabilities.

CASE STUDY: SYSTEM-BASED INTEGRATION

System-based integration depends on developing standardized data sources and software systems that support their integration. As in the previous case study, maintaining data quality is a primary issue. This demands standards for content and format be enforced at the level of the individual databases to be integrated. It also demands special attention to the fields in the databases that ultimately enable the linkages. Erroneous data in these fields can result in, at best missing data and at worst inclusion of inappropriate data in analyses.

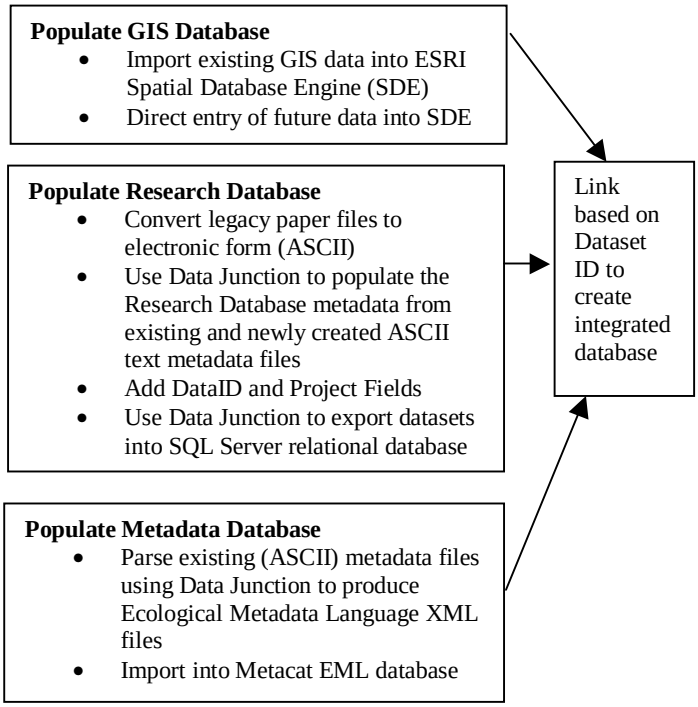
One example of a system-based integration project is underway at New Mexico State University (Figure 2). The Jornada Basin LTER (JRN) and the U.S. Department of Agriculture (USDA) Agricultural Research Service (ARS) Jornada Experimental Range (JER) are currently developing three separate databases that require integration to facilitate information management and improve discovery of and access to JRN and JER research datasets and associated metadata. The databases include a Research database for storing research datasets and documentation, a centralized Geographic Information System (GIS) database for storing GIS and Remote

Sensing data, and a metadata search engine (Metacat) database for querying metadata. These databases are being integrated to ensure that data and metadata within these applications are synchronized and accurate as well as to decrease redundant data storage as much as possible. Through integration of these databases, applications developed to support the automated Jornada Basin Information Management System (IMS) will be much more powerful and useful to information managers, researchers, and policy makers.

Populating the Research Database

All JRN research data and associated metadata are, and will be, archived as ASCII text files. These text files will be modified to permit the Research Database to be integrated with the GIS and Metacat databases. Modifications include adding project and dataset ID fields and ensuring that all data and metadata files conform to the standardized file formats. JER data and documentation will be converted to standard formatted ASCII text files as JER adopts the JRN IMS for archiving and managing research datasets and documentation. Some JER datasets date back to 1912 and many will need to be transcribed into electronic format.

Figure 2: Developing an Integrated GIS and Research Database



All archived JRN and JER text files will be parsed into a Research Database running Microsoft SQL 2000 Server. The entire process of parsing the text files to create or update the Research database will be scripted and automated using Data Junction Enterprise Integration Studio 7.51 (DJ) (Table 2).

Table 2: Software used in system-based integration

Software	Use
Data Junction Enterprise Integration Studio http://www.datajunction.com	Parsing of text documents into structured forms
ESRI ArcGIS http://www.esri.com	Input, retrieve, and manage spatial and remote sensing data and associated metadata stored in the Geographic Information System (GIS)
ESRI Spatial Database Engine (SDE) http://www.esri.com	Interface between relational database, GIS software and distributed applications
Microsoft SQL 2000 Server http://www.microsoft.com	Relational database used to support the Research Database and SDE
Metacat http://www.ecoinformatics.org	Customized metadatabase for use with EML-compliant metadata

Populating the GIS Database

The GIS database will provide centralized storage and access to spatial data such as remote sensing data covering the Jornada Basin or GPS data collected by JRN and JER researchers and technicians. Previously, all spatial data and metadata have been stored on the GIS Specialist's computer. Sharing spatial data and documentation in the past has been somewhat difficult. Now, by using newer spatial database software, ESRI ArcSDE 8.1 (SDE), it is much easier to share spatial information. The centralized GIS will be an integral part of the IMS; allowing spatially referenced querying of research data, remote sensing data, and associated metadata.

Existing remote sensing data imagery, coverages, and shapefiles will be imported into SDE using tools included in ESRI ArcGIS 8.1 software. Future GIS data and metadata collected will be stored and maintained in SDE.

Populating the Metacat Database

The Metacat database provides redundant storage of dataset documentation to allow the metadata to be queried using web-based applications [13]. Metacat uses the Ecological Markup Language (EML) to store metadata and for communication with the Metacat server using the XML protocol [14]. EML provides a standard exchange format for exchanging ecological metadata. EML will help reduce and spread the costs of application development by allowing development to be spread amongst the ecological research community. As new Metacat enhancements or EML based applications are released, they could be readily added to the IMS with minor modifications. The Jornada Basin Metacat server

will be linked to a LTER Network Office Metacat super node in the future. Metacat will provide another method for researchers to discover JRN and JER datasets; improving visibility and usefulness of JRN and JER scientific datasets.

Data Junction (DJ) will be used to create and populate EML compliant XML files, which subsequently can be imported into the Metacat database. DJ will also be used to automate this process in order to keep metadata stored in Metacat synchronized with metadata from the archival project and dataset documentation ASCII text files.

Application Development: The Need for Integrated Databases

The IMS will provide a web-based dynamic, interactive mapping and querying application that will be a powerful tool for JRN and JER information managers, researchers, and visitors. Some uses of the application include facilitation of research site selection and approval as well as automation of storage and retrieval of restricted and unrestricted spatial and non-spatial data and metadata into and from the IMS. Such an approach assists in land management decisions and eco-health evaluations, as well as monitoring of spatial and temporal vegetation changes. In order to achieve this level of functionality, the Research, GIS, and Metacat databases must be integrated, or linked.

Challenges: There are several limiting factors, or challenges, that had to be addressed in order to integrate the Research, GIS, and Metacat data to support the IMS, as well as develop the IMS. These challenges include adopting EML and Metacat, centralizing the GIS, and populating the Research database.

By adopting standards such as EML and utilizing Metacat software, it is hoped that the IMS can be enhanced in the future by using EML-based applications developed by other ecological research organizations. By developing the IMS to utilize EML, our development efforts can subsequently be used and enhanced by other organizations. Using tools such as DJ and XML Spy IDE Suite 4.1 for data conversion and extending the EML schema can greatly reduce development time and costs.

Linkages: Dataset IDs will be the common link that relates the Research, GIS, and Metacat databases. Other common thematic, temporal, and keyword fields will be created in the databases to allow queries of spatial and non-spatial data and associated metadata. If needed, pivot or lookup tables would be created in the databases using the common ID, thematic, temporal, and keyword fields. Stored views will be used to simplify development efforts where possible.

Steps for Integrating Research, GIS, and Metacat Data

The following list contains the steps required to integrate the Research, GIS, and Metacat data prior to the development of the IMS application:

1. Assign and add dataset IDs to dataset data and documentation files.
2. Assign project IDs and add dataset and project IDs to project documentation files.
3. Parse project and dataset data and documentation files to populate IMS database.
4. Perform QA/QC on parsed IMS database and correct any errors found in the archived ASCII text files and/or parsing scripts.
5. Parse project and dataset data and documentation files again if any errors were found during QA/QC of the IMS database.
6. Import or load existing spatial and remote sensing data and metadata into the GIS database.
7. Perform QA/QC on GIS data and metadata stored in the GIS database and correct any errors directly within the GIS database.
8. Add dataset IDs to research site location layers and features attribute tables stored in the GIS database.
9. Parse project and dataset documentation ASCII files and related IMS and GIS database tables to create EML XML DTD or schema files.
10. Perform QA/QC on parsed EML XML files and correct any errors found in the archived ASCII text files and/or parsing scripts.
11. Parse project and dataset data and documentation ASCII files and related GIS database tables again if any errors were found during QA/QC of the EML XML files.
12. Import EML XML files into the Metacat server database.
13. Import remote sensing images into the GIS database.

Upon completion of the sequence of steps, the IMS and GIS Internet and Intranet applications can then be developed and implemented.

The JRN and JER have completed the planning and evaluation stages of this IMS project. Data integration is currently underway. The project is designed to be modular to allow for prioritizing and planning the project development cycle.

RESEARCH AREAS FOR DATA INTEGRATION

The difficulty of a data integration project is directly proportional to the scale of the integration process. As the case studies above show, integration in specialized projects and systems can be accomplished using conventional software tools and information systems. However, for global-scale projects, which must integrate data from a huge number of data sources, these methods are not practical. The most common approach is to focus

on a few, large, very standardized, data sources. However, this approach excludes the vast majority of data sets, which are not standardized.

The challenges posed by dealing with large amounts of heterogeneous data are increasingly being confronted by developing techniques of data warehousing and data mining [15, 16]. Chen [10] calls for the application of data mining techniques to be used in conjunction with techniques for uncertain reasoning, such as fuzzy logic, genetic algorithms, Bayesian networks and rough set theory. Below is a brief outline of how a few of these techniques might be used in integrating ecological data.

Fuzzy Logic: Fuzzy logic allows probabilistic statements to be made about the true state of a variable. For example, for land cover classifications derived from remotely sensed data, you might conclude that there is a 75% chance that an area is forest and a 25% chance that the area is a shrubland. Traditional forms of analysis demand that we go with our best guess. However, with fuzzy logic, analyses can also consider additional guesses, each with its associated probability.

For ecological research, where qualitative determinations for land cover, habitat, community type and even taxonomic identity [11] are often suspect, (especially when integrating data from diverse sources) data integration incorporating fuzzy logic offers opportunities to incorporate a larger amount of information into analyses. Its use is not widespread in ecology, although it has been used in ecological applications in the context of remote sensing [17], ecological decision support [18] and modeling and prediction [19-21]. In our case studies, above, fuzzy logic could be applied to geographical locations that have varying degrees of specificity, and to land cover designations.

Evolutionary Algorithms: Evolutionary algorithms use a process of highly iterative trial and error to derive functional relationships and estimate parameters. Although primary uses have been primarily for developing search strategies and modeling, use of evolutionary algorithms holds promise for "harmonizing" data sources where the functional relationships between two ways of measuring an environmental variable are unclear. In our case studies, these techniques could be applied to harmonizing measurements taken at different scales or using different methodologies.

Data Mining Techniques: Traditional database approaches have had great difficulty dealing with heterogeneous information sources. However, techniques used in the rapidly evolving field of data mining can help to surmount these difficulties [15, 16]. Clustering, classification, and association rules have obvious uses in the data discovery process. However, they can also be used for at least partial automation of data integration by helping to identify similar variables in different datasets. Some of these techniques are widely used in ecology, although typical uses are more oriented towards data analysis than data integration.

Meta-Analysis: In addition to the techniques listed by Chen [10], meta analysis provides tools for a different approach to integration. Ecological meta-analysis integrates results from previously published studies to attack broader questions and to strengthen individual conclusions [22]. The effect sizes observed from a variety of studies, each using different data sources and methods can be statistically combined to reach new conclusions.

CONCLUSIONS

Use of advanced data mining and techniques for dealing with uncertainties would be a powerful approach for the facilitation of ecological synthesis, but such developments need to be coupled with specific enhancements that ensure use by the scientific community. As previously noted, scientists place a high value on data quality. Unlike some disciplines where a final product is evaluated on its intrinsic merits (regardless of origin), scientific research products are evaluated based primarily on the methods and data used to produce them. Complex, integrated datasets pose problems because a full explanation of data sources may be impossible and reviewers and readers need assurance that results are real, and not artifacts of the integration process.

Enhanced analysis systems are required to support use of integrated data. These systems need to make datasets auditable, so that each datum used in an analysis can be traced back to its original source and transformations reproduced. Reproducibility is critical to developing trust in a scientific product.

Second, tools need to be developed that facilitate sensitivity analyses, wherein specific data sources can be added or subtracted from an analysis. This allows researchers to determine whether particular data sources have undue influence on the final result, or whether their conclusions are robust with respect to changes in data sources.

Finally, visualization techniques can clarify the roles of individual data sources. Animated graphs which highlight specific data sources make it possible to review a large number of data sources in a short period of time. This approach assures that patterns are discernable within, as well as between, data sources.

Meeting the challenges inherent in large-scale data integration is the subject of ongoing research in both the computer science and ecological communities.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grants No. DEB-0080381 and DEB-0080412. William Michener, James Laundre and Karen Baker provided constructive comments on the manuscript.

REFERENCES

[1] National Research Council, Bits of Power: Issues in Global Access to Scientific Data. National Academy Press, Washington, D.C., 1997.

[2] J.H. Porter and J.T. Callahan, Circumventing a Dilemma: Historical Approaches to Data Sharing in Ecological Research. in W.K. Michener, S. Stafford and J.W. Brunt eds. Environmental Information Management, Taylor and Francis, Bristol, PA, 1994, 193-203.

[3] D.E. Strebel, B.W. Meeson and A.K. Nelson, Scientific Information Systems: A Conceptual Framework. in W.K. Michener, S. Stafford and J.W. Brunt eds. Environmental Information Management, Taylor and Francis, Bristol, PA, 1994, 59-85.

[4] W.K. Michener, J.W. Brunt, J.J. Helly, T.B. Kirchner and S.G. Stafford, "Non-Geospatial Metadata for the Ecological Sciences". Ecological Applications, Vol. 7, No. 1, 1997, pp. 330-342.

[5] B.R. Schatz and J.B. Hardin, "NCSA Mosaic and the World Wide Web: Global Hypermedia Protocols for the Internet". Science, Vol. 265, 1994, pp. 895-901.

[6] B. Hanson and R. Coontz, "A Computer Science Odyssey". Science, Vol. 293, No. 5537, 2001, pp. 2021.

[7] J. Pfaltz, Differences between Commercial and Scientific Data. in Scientific Database Management, a Report to the National Science Foundation, 1990.

[8] R.J. Robbins, "An Information Infrastructure for the Human Genome Project". IEEE Engineering in Medicine and Biology, Vol. 14, No. 6, 1995, pp. 746-759.

[9] J.H. Porter, Scientific Databases. in W.K. Michener and J. Brunt. eds. Ecological Data: Design, Processing and Management, Blackwell Science Ltd., London, UK, 2000.

[10] Z. Chen, Data Mining and Uncertain Reasoning: An Integrated Approach. John Wiley and Sons, New York, 2001.

[11] D. Maier, E. Landis, J. Cushing, A. Frondorf, A. Silberschatz, M. Frame and J.L. Schnase. Biodiversity and Ecosystem Informatics: Report of an NSF, USGS, NASA Workshop on Biodiversity and Ecosystem Informatics Held at NASA Goddard Space Flight Center, June 22 - 23, 2000., NASA, Beltsville, MD, 2001, 36.

[12] B.P. Hayden, R.D. Dueser, J.T. Callahan and H.H. Shugart, "Long-Term Research at the Virginia Coast Reserve: Modeling a Highly Dynamic Environment". Bioscience, Vol. 41, 1991, pp. 310-318.

[13] M.B. Jones, C. Berkley, J. Bojilova and M. Schildhauer, "Managing Scientific Metadata". IEEE Internet Computing, Vol. 5, No. 5, 2001, pp. 59-68.

[14] R. Nottrott, M.B. Jones and M.P. Schildhauer, Using XML-Structured Metadata to Automate Quality Assurance Processing for Ecological Data. in Proceedings of the Third IEEE Computer Society Metadata Conference, (Bethesda, MD., 1999), IEEE.

[15] M.J.A. Berry and G.S. Linoff, Mastering Data Mining: The Art and Science of Customer Relationship Management. John Wiley & Sons, Inc., New York, 2000.

[16] M.A. Bramer, Knowledge Discovery and Data Mining. The Institution of Electrical Engineers, Herts, UK, 1999.

[17] G. Metternicht, "Assessing Temporal and Spatial Changes of Salinity Using Fuzzy Logic, Remote Sensing and GIS. Foundations of an Expert System". Ecological Modelling, Vol. 144, No. 2-3, 2001, pp. 163-179.

[18] D.M. Stoms, J.M. McDonald and F.W. Davis, "Fuzzy Assessment of Land Suitability for Scientific Research Reserves". Environmental Management, Vol. 29, No. 4, 2002, pp. 545-558.

[19] A. Pistocchi, L. Luzi and P. Napolitano, "The Use of Predictive Modeling Techniques for Optimal Exploitation of Spatial Databases: A Case Study in Landslide Hazard Mapping with Expert System-Like Methods". Environmental Geology, Vol. 41, No. 7, 2002, pp. 765-775.

[20] L.O. Odhiambo, R.E. Yoder, D.C. Yoder and J.W. Hines, "Optimization of Fuzzy Evapotranspiration Model through Neural Training with Input-Output Examples". Transactions of the ASAE, Vol. 44, No. 6, 2001, pp. 1625-1633.

[21] V. Gomez and A. Casanovas, "Fuzzy Logic and Meteorological Variables: A Case Study of Solar Irradiance". Fuzzy Sets and Systems, Vol. 126, No. 1, 2002, pp. 121-128.

[22] C.W. Osenberg, O. Sarnelle, S.D. Cooper and R.D. Holt, "Resolving Ecological Questions through Meta-Analysis: Goals, Metrics, and Models". Ecology, Vol. 80, No. 4, 1999, pp. 1105–1117.