01001100 01010100 01000101 01010010

LTER DataBits

Information Management Newsletter of
The Long Term Ecological Research Network

01001100 01010100 01000101 01010010

Feature Articles

News Bits

Good Reads

Frequently Asked Questions

Calendar

DataBits: An electronic newsletter for Information Managers. ----- *Spring 2001*
http://www.lternet.edu/documents/Newsletters/DataBits/01spring

Featured in this issue:
A Custom data management system, International LTER Information Management Workshop, The NASA Scientific Data Purchase Program, and Web-based data entry for dummies.

DataBits continues as a semi-annual electronic publication of the Long Term Ecological Research Network. It is designed to provide a timely, online resource for research information managers and to incorporate rotating co-editorship. Availability is through web browsing as well as hardcopy output. LTER mail list IMplus will receive DataBits publication notification. Others may subscribe by sending email to majordomo@lternet.edu with two lines "subscribe databits" and "end" as the message body. To communicate suggestions, articles, and/or interest in co-editing, send email to databits-ed@lternet.edu.

----- *Co-editors: Brent Brock (Konza Prairie) and Wade Sheldon (Georgia Coastal Ecosystems)*

Feature Articles

A Standard for Creating Dynamic, Self-documenting Tabular Data Sets Using Matlab®

-Wade Sheldon, Georgia Coastal Ecosystems (GCE)

One of the biggest challenges I have faced setting up the GCE data management system is providing support for online data analysis and flexible formatting for all data sets. When I asked our site scientists to identify features that would enhance their ability to use shared project data, the most common responses were: 1) online plotting to preview data, 2) support for sub-sampling data sets (i.e. downloading only portions of interest), 3) multiple file format options to minimize post-download processing. Clearly, developing effective protocols for storing dynamic, computer-readable data sets needed to be a high priority in our information system.

One approach to this problem would be to store all data in a relational database management system and provide plotting, querying and formatting capabilities through server- and client-side web applications. While this is certainly an effective approach for large homogenous data sets, like those from our core monitoring efforts, it has a number of drawbacks for managing data from individual studies. Some common study elements, such as replication and repeated measures, are difficult to accommodate in relational models. Also, the expertise and administration required to maintain highly diverse data sets (i.e. "wide" databases) in a RDMS and assist users with queries might become a burden for data management staff (Porter, 2000).

Given these limitations, we decided to develop a custom software solution to process, store, analyze, display and format tabular data sets from research studies. The software and storage specification were developed using Matlab® (The Mathworks Inc., http://www.mathworks.com), an open-source, multi-platform programming language for numerical analysis and data visualization. Matlab is a dominant programming language in oceanographic research and is used by many GCE investigators, so we felt this tool provided the best potential for long-term code support and collateral usage. In addition, the availability of add-on function libraries ("Toolboxes") for accessing Matlab programs from web forms, connecting to databases, and displaying geocoded data on map projections will allow us to meet all our analysis and display needs with a common set of tools.

The initial results of this effort are now complete and in use at our site, as described in our information system guide (see references). The primary components are: 1) a standard for storing fully-documented tabular data sets as Matlab data structures, and 2) a set of functions constituting the 'GCE Data Toolbox'. Data structures are multi-dimensional arrays organized into named fields, each of which can store data of any size and complexity (unlike conventional database fields). This capability was exploited to encapsulate variable amounts of structured metadata inside a single structure field, allowing the metadata to be tightly coupled to the data set. Other fields contain column descriptor information, such as the names, descriptions, units, precisions, data storage types, logical variable types (i.e. domains), and numerical characteristics of each data column. The data set itself is stored as an array with each column containing either floating-point, integer or string (mixed alphanumeric) values. A matching string matrix for storing QA/QC flags is also supported, allowing flexible handling of flag information separately from the data values.

The functions in the GCE Data Toolbox provide a layer of abstraction for users, allowing data structures to be created, manipulated, analyzed, and exported using simple commands without knowing anything about their actual composition. The tools are also 'data aware', in that they use the metadata information stored in the structure to validate the data set and apply a semantic approach when performing statistical analyses. For example, the column statistics function will calculate a median but not a mean on an integer column, and will compute a vector average rather than an arithmetic average if a floating-point column has a numeric type of 'angular'. Another powerful feature of these tools is that they transparently store processing history information with each function iteration and update column descriptor metadata fields to reflect the actual composition of the data structure at the time the metadata is parsed to generate the data documentation. Providing automatic linkages between data and metadata to maintain the quality and validity of data sets has been a major design goal of the GCE Data Toolbox.

Various analytical functions have also been developed to provide basic database functionality for data structures. These functions support column selection, multi-column bi-directional sorting, natural language multi-column queries to select rows, and multi-column aggregation (with statistical analysis in each aggregate for specified columns). This latter capability has already proven to be a valuable research tool, allowing us to sort and aggregate large plant monitoring data sets by various categorical variables, quickly producing summary statistics for various levels of detail in the study. Work is now underway to develop WWW and Matlab GUI interfaces to allow non-Matlab users to analyze data sets online as well as offline.

Developing custom software to manage data is certainly not appropriate for every situation, but it does offer unique opportunities to solve scientific problems not easily addressed using business-oriented tools. I will report back on our progress as we continue to develop this technology and put it to task at our site.

_____

References

GCE LTER Information System Guide: http://gce-lter.marsci.uga.edu/lter/research/guide/gce-is.htm

Porter, J.H. 2000. Scientific Databases. In: W.K. Michener and J.W. Brunt (Editors), Ecological Data - Design, Management and Processing. Methods in Ecology. Blackwell Science Ltd., London, pp. 48-69.

International LTER Information Management Workshop

- John Porter (VCR)

Köszönöm! There, now that I have dispensed with my knowledge of the Hungarian language (it means “Thanks!”) I can tell you about the Information Management Workshop Kristin Vanderbilt (SEV) organized along with Peter McCartney (CAP) and myself. Our local host was Dr. Edit Kovácsné Láng who did an excellent job in providing us housing and food at the “Weekend Panzio” (“Panzio” is Hungarian for “hotel”) and excellent computational and network facilities at the nearby Botanical Institute. The workshop was aimed at providing some training in information management techniques for participants from the Czech Republic, Hungary, Poland, Romania and Slovakia. Interestingly, discussion groups incorporating members from different countries all needed to use English, as it was the only language spoken in common by the entire group.

During five intensive days (8 AM until 6 PM each day), the workshop covered a wide variety of topics (Concepts in Ecological Information Management, Distinctive Characteristics of Ecological Information, Developing Good Collaborative Relationships Between Scientists and Information Managers, Information Management Policies, Using Microsoft ACCESS, Developing World-Wide Web Pages, Database Design and Modeling, The Data Cycle, Quality Control and Quality Assurance, Techniques for Connecting Databases to the WWW, Administration of WWW servers, Ecological Metadata and the Global Terrestrial Observing System [GTOS]) in lecture, discussion and laboratory formats. The text for the course was the new “Ecological Data: Design, Management and Processing” book, edited by William Michener and James Brunt of the LTER Network Office. Material in the form of Powerpoint slides were also obtained from William Michener of the LTER Network Office and Ray McCord and Dick Olson of Oak Ridge National Laboratories. During laboratory exercises, workshop participants designed databases, practiced simple QA/QC procedures, ran database queries and created web pages.

However not everything was work! We discovered the pleasures of sausage, cheese, rolls and paprika (mild green peppers) for breakfast, and a variety of meat and cheese dishes for lunch and dinner. It was also a good opportunity to learn about the challenges faced by information managers at LTER sites in Eastern Europe. Not surprisingly, the idea of integrated, project-based information management was relatively new to many of the participants. Similar to the U.S. when LTER was initially started, there is no existing pool of information managers. They also face the challenge, similar to the U.S., that there is not a clear “career path” for information managers. They must also contend with having fewer resources than those provided by the National Science Foundation for U.S. LTER sites.

Participants in the workshop set up a home page containing all the workshop training materials at: http://www.krnap.cz/lter/, and photos of the workshop are available at: http://www.vcrlter.virginia.edu/images/lter_network/ILTER_Hungary_2000/IM_Training/

Following the workshop, Kristin spent an additional several weeks in Hungary helping to establish soils research plots at the Síkfökút deciduous forest LTER site near Eger, Hungary in association with Dr. János Attila Tóth. I stayed an additional week and made one-day visits to each of the Hungarian LTER sites. Edit andher colleagues at the Kiskun LTER site provided a deluxe tour of the sand dunes, grassland and shrublands at their site. János, along with Kristin, provided a tour of the Síkfökút deciduous forest site (like many U.S. LTER sites, it is an old IBP site). I also drove down and circled Lake Balaton, a large (70 km long) clear-water lake. As with U.S. LTER sites, there were interesting ecological challenges associated with each of the sites. The Kiskun site has experienced fires, which are a novel phenomenon in a generally human-dominated landscape. At Síkfökút, one of the two main oak species has experienced a major die-off in the last decade. Lake Balaton is a major tourist destination, with all the challenges that brings. Photos from each of the sites are available at: http://www.vcrlter.Virginia.EDU/images/lter_network/ILTER_Hungary_2000/ and detailed videos of the tours are available on request.

News Bits

NASA Data Purchase Program Offers No-cost Acquisition of Remote Sensing Data

- Greg Hoch, Konza Prairie

The Scientific Data Purchase (SDP) program was initiated in 1997 and managed by NASA's Commercial Remote Sensing Program (CRSP) at the Stennis Space Center. Through this program NASA purchases image data from 5 private contractors and distributes the data to it's users. These data range from historic ortho-rectified Landsat products, radar (IFSARE) data, and 1 to 4 m multi-spectral imagery. The goal of this program is to provide NASA affiliated researchers with imagery to address questions involving land cover and land cover change, interannual climate variability, and long-term climate change. The website for ordering data (http://www.crsp.ssc.nasa.gov/scripts/datapurchase/dbmain.asp) requires that you register but is relatively user friendly. Data requests are submitted 4 times a year and reviewed by committee. The status of your request can be checked on the SDP website.

To date, Konza Prairie has acquired 4 IKONOS images from the 2000 growing season as well as a 4 m resolution Digital Elevation Model (DEM) generated from stereo-pairs of IKONOS imagery. These data are currently being used to map the locations of shrub islands on Konza and relate the spatial patterns of these islands to burn frequency. Konza also has 5 Landsat and 2 SPOT images from summer 2000 that were acquired independently and we are currently comparing a suite of spatial statistics for individual watersheds from these sensors across a range of pixel sizes, or spatial resolutions. We are also using the IKONOS imagery to examine spatial and temporal patterns related to grazing from our bison herd. The DEM is being used for a host of geomorphology as well as ecological projects. The NASA SDP program has provided over $45,000 worth of remote sensing data to Konza Prairie so far and has significantly enhanced our

research program.

Simple Tools for Creating Web-based Data Entry

Brent L. Brock, Konza Prairie

Web-based data entry, query, and retrieval are useful for a variety of functions. However, development of such databases can be time consuming and intimidating to those lacking adequate programming or scripting skills. We recently developed a simple Web-based data entry and retrieval prototype for the Konza Environmental Education Program using database connectivity tools available in Microsoft FrontPage 2000 and FrontPage server extensions. This prototype currently allows students to enter field data collected at Konza Prairie Biological Station into a centralized database using a standard Web browser. The data is automatically combined with data collected by students at other schools for further analysis.

The FrontPage tools are relatively easy to use despite vague documentation. The first step is to create a form for data entry and save it in an active server page.FrontPage has a comprehensive toolkit for form design and allows the user to set ranges and rules for data fields for on-the-fly integrity checking. The next step is to assign a database connection and table for data storage. FrontPage has many options for form handling and can even create a new database connection and table that stores the data in an Access database within the current Web. For our prototype we created an ODBC system database with multiple tables so that each data entry form has its own table within the Access database. These tools allow individuals with moderate computer skills to relatively easily create web-based data entry forms and databases without any SQL or scripting skills. In our current prototype data query is limited to retrieving a table of all data contained in a database table. However, we will soon implement more sophisticated query tools that will allow students to view summary results of their research projects with a click of a button.

Good Reads

Good Reads:

Public Access and Use of Electronically Archived Data: Ethical Considerations

- Brent L. Brock, Konza Prairie

Davis, M.A., D. Tilman, S.E. Hobbie, C.L. Lehman, P.B. Reich, J.M. Knops, S. Naeem, M.E. Ritchie, and D.A. Wedin. 2001. Public Access and Use of Electronically Archived Data: Ethical Considerations. Bulletin of the Ecological Society of America 82:90-91.

This brief article explores the ethical considerations regarding publishing and use of publicly available archived data. The authors stress the need for adoption of a code of ethics to encourage data sharing and protect researchers from inappropriate or unethical use of their data. This article provides food for thought for data managers and ecology researchers alike.

Evolution of a Multisite Network Information System: The LTER Information Management Paradigm

- Wade Sheldon (GCE)

Baker, K.S., B.J. Benson, D.L. Henshaw, D. Blodgett, J.H. Porter, and S.G. Stafford. 2000. Evolution of a Multisite Network Information System: The LTER Information Management Paradigm. BioScience, 50(11): 963-978.

Once again we have the opportunity to use this space to showcase an outstanding contribution to the field of Ecological informatics by members of our own inner circle. Like 'Ecological Data', reviewed in the Fall 2000 issue of Databits, this paper encapsulates years of LTER information management knowledge and experience in one source. The central theme of the paper is the history and ongoing development of the LTER NIS, presented as a review and case study for application in other research areas. The authors also use the article to educate readers about the basic tenets of scientific information management by providing glossaries of terminology, conceptual diagrams, and lists of core design elements for developing a NIS. This is an inspiring article which should be on every data managers reading list, and should be required reading for any class on scientific information management.

FAQs: Frequently Asked Questions

Where can I get information about metadata and other issues in ecoinformatics?

Two excellent sources of information for ecoinformatics are Ecoinformatics.org (http://www.ecoinformatics.org/) and the LTER Metadata Task Group site (http://caplter.asu.edu/data/metadata/). These sites contain a wealth of information about metadata standards, tools, and other information regarding ecoinformatics.

Calendar

01Jan LTER NAB preparation meeting
30-31Jan LTER mini-NIS meeting
24-28Feb SDSC NPACI All-Hands Meeting
01Apr LTER/KDI? Metadata meeting
25Apr LTER CC Meeting
14-16 June Data Synthesis workshop
01Aug LTER DM meeting
01Aug ESA meeting
05Aug Scalable information systems workshop

Highlights from Mini-NIS meeting **16 Feb 2001**

Dates: 30-31 January 2001
Location: LTER Network Office
From: K.Baker, J.Brunt, D.Henshaw
Highlights posted at: http://www.fsl.orst.edu/lter/im/nis_mtg_2001.htm

Present at the meeting:
Karen Baker (PAL/SiteDB module leader)
James Brunt (NET/ASBIB&Personnel module leader)
Don Henshaw (AND/Climdb module leader)
David Blankman (NET)
Richard Dahringer (NET)
Troy Maddux (NET)
Bill Michener (NET)
Patricia Sprott (NET)

This meeting was originally proposed as a post-Snowbird meeting by Baker and Henshaw, but was instead supported by the Network Office as part of the on-going IM mission.

NIS:
Karen, Don, and James will report to IMEXEC the need for a revised NIS plan and NIS working group re-activation. The NIS working group has been without a chair since 1997. This meeting functioned as a NIS working group meeting by having the leaders of two of the central NIS modules present. Although not the original focus of the meeting, the work with ClimDB and SiteDB quickly involved the broader issues of definition and integration of modules into a Network Information System.

SiteDB:
Background: The SiteDB module, designed and functioning since Aug 2000, was implemented originally at Palmer site using SQLServer with a web interface scripted in Perl. The module performs four functions:

input, view, modify and compare site data. Code was delivered to Network Office in August 2000 and documentation in September. The code has been migrated to SQLServer and subsequently to an Oracle implementation at the Network Office. Troy Maddux had populated the site table with information from site web pages and from the existing network collection of static site descriptions. David Blankman and James had created a preliminary database design integrating site administrative information, site physical description information, and personnel from SiteDB with network personnel and publication tables.

At this NIS meeting, a demo featured the current Network Office implementation of the input and view functions of SiteDB. After demos and discussions, a new and more normalized module design emerged permitting multiple metadata types for research sites and locations and allowing hierarchical referencing of sites and sub-sites. This extensible design will facilitate inclusion of site-level metadata for ClimDB as well as for any data module required of NIS. The design modifications also ensure ease of accommodation of research sites other than LTER sites such as OBFS and Forest Service. This integrated model forms the core of the research support or administrative side of the NIS.

David will make the necessary changes to the data model and will provide an updated Erwin Studio design (ERD) for posting by 9 Feb. (This is now available at http://sql.lternet.edu/lternis/). David will also implement the tables in the database and will post definitions for table fields. At that point Troy and James will begin the task of web interface integration and population. Deadline for this work is March 1.

ClimDB:
The ClimDB data model and interface have been designed to be extensible. Don Henshaw implemented ClimDB at the AND site and extended the concepts to include a new module, HydroDB. Initial tests and demonstration indicate that the ClimDB data model and interface are extensible for selected other data types without requiring programmatic changes.

Don will initiate remote harvest tests in order to include all sites previously part of the database, as well as encourage participation from all other sites. Draft documentation of ClimDB programs and processes will be completed in February. Climdb is undergoing some minor programming changes to make it more robust. Primary emphasis will be the refinement of the administrative front-end, in particular, the improvement of the harvester to include a user-feedback mechanism. Consideration will also be given to improving the query interface and graphical displays. Richard and James will complete these program changes by 15 March with complete harvest of all site data expected at this time. With the extensions mentioned above in SiteDB, ClimDB forms the core of the research side of the NIS.

Complete notes of our ClimDB discussion are available: http://www.fsl.orst.edu/lter/im/climdb_notes_01.htm

HydroDB:
HydroDB is a project being led by Don Henshaw in coordination with the USFS Forest Health Monitoring (FHM) program to share intersite hydrology data similarly to LTER climate data using the ClimDB model. Don was able to incorporate hydrology data directly into ClimDB without additional programming. This bodes well for using this prototype to produce other value-added cross-site databases. This is exciting news.

All-Site Bibliography:
The group developed a plan for revision and enhancement of functionality for the all-site bibliography. The ASBIB database was last updated in 1998 and has not been updated since because of the difficulty in working with the ever-evolving site-specific scripts that are required to integrate the data. The group present agreed that an enabling solution would be to consider implementation of the LTER Network bibliographic

record standard within the powerful new Endnote 4.0 software package available for a variety of computer platforms. After consulting with the Executive Director, James Brunt will investigate purchase of this software for those sites that would like to try it as an alternative for providing their site bibliography data to the NIS. James would also put together the necessary input and output filters to facilitate updates. Sites would have the option of producing a standardized bibliographic format file independently, or using the Endnote software to do it. The FREE software would provide a carrot for getting the data renewed on a more regular basis and would allow the Network Office to use the ISI Reference Web Poster as an end-user interface which provides a great deal of the desired searching and sub-setting functionality.

All-Site Personnel:
The personnel table will interface with SiteDB and all NIS modules, and personnel role types have been redefined.

Advancing the Sharing and Synthesis of Ecological Data: Guidelines for Data Sharing and Integration

Working Group 1: Partnerships for Inter-site Synthetic Research and Initiation of Standards Development

Location: Sevilleta LTER Field Station
Dates: June 14-16, 2001
Organizers: Barbara Benson, Tim Fahey, Alan Knapp, and Dick Olson

Last summer at the LTER All Scientists Meeting (ASM), the LTER Information Managers sponsored a workshop entitled “The Partnership between Long-term Ecological Research and Information Management: Successes and Challenges”. This forum generated a productive dialogue among scientific researchers and information managers by examining partnerships between researchers and information managers both at individual sites and for inter-site research. A follow-up activity was funded by the LTER Network Office to support meetings of two working groups this year to more strongly solidify the partnerships between scientists engaged in long-term ecological studies and scientists involved with information management. We will focus on particularly salient issues related to data standardization efforts, data quality assurance and quality control, and metadata.

A major focus of the working groups will be to apply the results of their analysis to a specific data collection and synthesis activity case study. As a case study, the working group will review and propose potential schema for standardizing measures of net primary productivity (NPP) as was discussed for the LTER-wide focus (ASM Workshop: Primary Productivity in Forests: Status of research and planning for the future, T. Fahey, HBR) and as an ILTER project (ASM Workshop: Biodiversity and net primary productivity demonstration projects for ILTER and GTOS, J. Gosz, SEV).

The first working group will meet June 14-16 and has set the following goals for the meeting:

1.Review synthesis activities (partnerships, draft paper from ASM Partnerships workshop, and ASM NPP activity and Science paper, etc).
2.Explore the development of standardized data collection protocols and other guidelines to compile data

with the intent of using such data in synthesis activities.
3.Review the use of methods to harmonize long-term data assembled for data synthesis activities.
4.Develop a plan for the development and synthesis of NPP data from the LTER , ILTER, and GTOS networks
5.Evaluate the barriers for acceptance and use of guidelines for metadata and system documentation.
6.Compare sources of variability within collections of data, including biome, interannual, methods, and sampling to determine if reducing variability associated with methods will improve the overall analysis and interpretation.
7.Develop a plan for workshop 2's (*QA/QC and Continuation of Standards Development)* focus on assessing the potential for developing and implementing standard procedures for long-term data QA/QC.

Ecological Society of America Workshop on "Scalable information systems: from laboratories to NEONs"

Location: Madison Wisconsin
Dates: Aug 5, 2001
Organizers: William Michener and Art McKee

W. Michener, LTER Network Office

Register early! (attendance is limited to 100; go to http://esa.sdsc.edu/) and come join the fun at the Year 2001 Ecological Society of America Annual Meeting in Madison, Wisconsin. The ESA meeting will be held at the Monona Terrace Convention Center, August 5 thru 10, 2001. As part of this meeting, a one-day workshop addressing computing, communications, data and information management at field stations will be held on Sunday August 5th. The workshop is being sponsored by the Long Term Studies Section of ESA, the Organization of Biological Field Stations, the San Diego Supercomputer Center, and the LTER Network Office.

The workshop has been organized by William Michener of the LTER Network Office and Art McKee of the H. J. Andrews Experimental Forest. Workshop speakers and titles include:

a. James Brunt (University of New Mexico) – *Computing environments*, *communications and networking*

b. Mark Schildauer (University of California – Santa Barbara) – *Metadata*

c. Dick Olson (Oak Ridge National Laboratory) – *Archives and regional databases*

d. John Porter (University of Virginia) – *Database approaches*

e. William Michener (University of New Mexico) – *QA/QC*

f. Cherri Pancake (Oregon State University) -- *Web interfaces and data & information portals*

g. Matt Jones (University of California – Santa Barbara) -- *Tools for data integration, analysis and synthesis*

h. Warren Cohen (Oregon State University) – *Approaches for scaling from the site to broader scales*

i. Robert Peet (UNC-Chapel Hill) – *Taxonomic and museum databases*

Interest in data- and information-related technologies has intensified in response to proposed plans for building a National Ecological Observatory Network. Despite interest in relevant technologies, most ecologists have not been able to keep pace with the rapid advances in computing, communications, and information management and analysis. This workshop is specifically designed to make ecologists aware of new and appropriate information technologies as they consider upgrading individual laboratories or expanding field station capabilities to those needed for NEON.

Each speaker will present a broad overview of their topic in a 30-45 minute period. Specific material covered by each speaker will include basic to advanced approaches that can meet needs ranging from those of individual scientists to reasonably well-equipped field stations that are contemplating expansion to a more sophisticated observatory. For each topic, a range of possible solutions, as well as costs, personnel requirements, and other factors will be discussed. Reference materials will be provided to all participants and approximately one-fourth of the workshop agenda will be set aside for discussion and question/answer sessions.

Monona Terrace Convention Center (foreground) in Madison, Wisconsin (from ESA web site).