LTER

Long Term Ecological Research

01001100 01010100 01000101 01010010

LTER DataBits

Information Management Newsletter of The Long Term Ecological Research Network

01001100 01010100 01000101 01010010

Feature Articles

Four part Semantics Series:

Editorials

News Bits

Good Reads

Frequently Asked Questions

DataBits: An electronic newsletter for Information Managers ----- Spring 2006 Issue (http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/06spring/)

Featured in this issue:

Welcome to a robust issue of the Spring 2006 DataBits'!!! We had many submittals this issue and the articles really show the diversity of this group. We hope you have as much fun reading this issue as we did putting it together. As many of the LTER sites have EML documents being generated, the focus of the IM community has shifted from generating EML to working with EML. There is a series of four articles in this issue of DataBits' that explore multiple approaches with differing timeframes that IM's are developing that will lead to better data discovery within EML. TRENDS also stops by to give an update on their progress. As data mangers, we often over-look some of the ways that we can outreach to our surrounding communities, and Chris Gardner from McMurdo Dry Valleys LTER provides a great avenue with the Children's Book Series. A pair of editorials consider MySQL and PostgreSQL as well as the usual collection of informative articles, good reads, and news bits. So sit back, grab a cup of coffee, and enjoy this issue of Databits'

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: Brian Riordan (BNZ), Steve Bauer (CDR)

Feature Articles

Mining and Integrating Data from ClimDB and USGS using the GCE Data Toolbox

- Wade Sheldon (GCE)

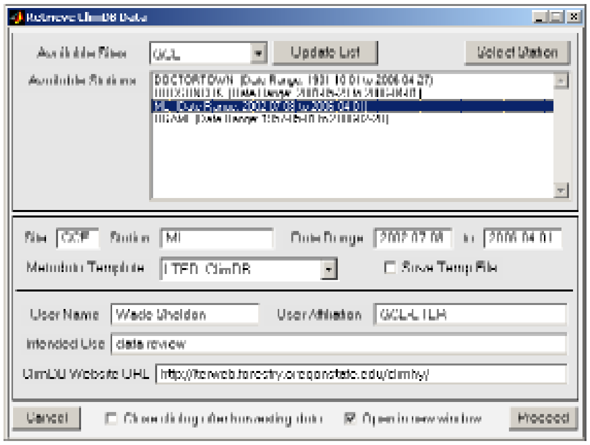
Introduction

Climate and hydrographic data are critically important for most long-term ecology studies, so integrated climate and hydrography databases such as ClimDB/HydroDB and the USGS National Water Information System represent major resources for ecologists. Both ClimDB/HydroDB and USGS also have web-based query interfaces, which provide convenient access to data and basic plots from many monitoring stations across the country. These centralized databases therefore significantly aid users with the first two phases of any data synthesis project: data discovery and data access.

Data synthesis doesn't stop with the initial download, though, and many users I've worked with quickly become frustrated performing the remaining steps that are typically required. For example, common follow-up tasks include parsing and importing data into spreadsheets or analytical software, assigning or harmonizing attribute names and units, and integrating data from multiple stations for comparative analysis. Automating these operations is highly desirable, but usually requires custom programming and is not practical for most researchers. Consequently some students and researchers avoid data synthesis all together, viewing it as either too tedious or difficult, while others request help with synthesis tasks from information management staff, adding to their workload.

As I've described in several prior Data Bits articles (1,2,3), at GCE we have developed a flexible environment for metadata-based data transformation, quality control, analysis and integration using the multi-platform MATLAB programming language (i.e. GCE Data Toolbox). This software was also used to develop an automated USGS data harvesting service for HydroDB that contributes near-real-time hydrographic data on behalf of 10 LTER sites to the ClimDB/HydroDB database on a weekly basis (4). In the remainder of this article I describe new data mining features recently added to this toolbox that allow users to interactively retrieve data from any station in ClimDB/HydroDB or the USGS NWIS (using MATLAB 6.5 or higher), and then easily transform and integrate these data sets to perform synthesis on their own.

Mining ClimDB/HydroDB



Retrieve ClimDB Data

Available Sites: GCE

Available Stations:
DOCTORTOWN (Date Range: 1931 10 01 to 2006 04 27)
[illegible]
ML (Date Range: 2002 07 08 to 2006 04 01)
[illegible]

Site: GCE Station: ML Date Range: 2002 07 08 to 2006 04 01

Metadata Template: LTER_ClimDB □ Save Temp File

User Name: Wade Sheldon User Affiliation: GCE-LTER

Intended Use: data review

ClimDB Website URL: http://fsweb.forestry.oregonstate.edu/climhy/

□ Close dialog after harvesting data ☑ Open in new window

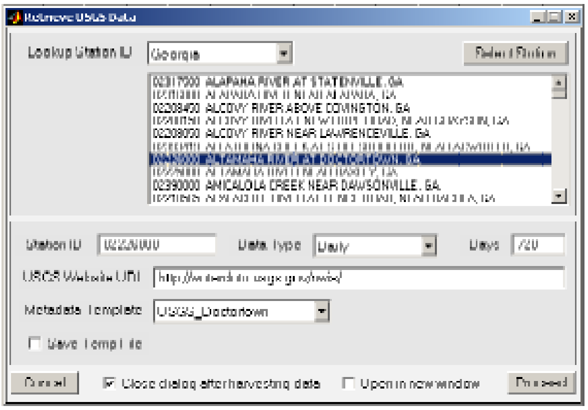
fig 1. ClimDB/HydroDB data mining dialog
(click on image for larger version)

Retrieving data using the ClimDB/HydroDB web application is a multi-step process, starting with site selection and user registration, followed by station and parameter selection, then specification of download options. In order to automate this process on behalf of users a complete "variables" report is initially retrieved from the ClimDB web server in XML format. This report is then parsed into delimited text format using XSLT and cached as a data table to provide an up-to-date list of stations and available parameters, which can be updated at any time on demand. This parameter table is used to populate controls on a GUI form (fig 1), which also prompts for and caches user name and affiliation information to pass to ClimDB along with each data request.

When the user selects a site using the drop-down menu at the top of the dialog, all available stations for that site are automatically displayed along with the temporal coverage for each. Pressing the "Select Station" button (or double clicking on the entry) copies the relevant site and station information to the data request form and fills in the complete date range, which the user can then fine-tune to cover only the date range of interest. Pressing the "Proceed" button sends a series of requests to the ClimDB web server via HTTP POST commands (with progress displayed after each step), until the requested file is transmitted and saved locally. The retrieved file is then parsed and transformed into a GCE Toolbox-compatible file format and loaded into the toolbox data set editor window for analysis, transformation and display (see http://gce-lter.marsci.uga.edu/lter/research/tools/toolbox_applications.htm). At present, daily data are requested for all available parameters, relying on GCE toolbox functions for data sub-setting and temporal re-sampling, but other options may be added in the future.

On the surface, the work flow just described only represents a convenient way to query ClimDB/HydroDB using MATLAB; however, several important transformations occur automatically during the data import step that can significantly streamline synthesis. For example, the GCE toolbox supports the use of user-defined metadata templates, which can be applied whenever data are imported from a particular source. These templates contain boilerplate documentation metadata as well as a table of attribute descriptors and QA/QC flagging rules for every variable that may be present in the source data (i.e. matched by field name). A default ClimDB template is included with the toolbox distribution, but users can edit this template in a GUI dialog and specify their own site-specific names and descriptions for ClimDB attributes (as well as additional QA/QC criteria), causing ClimDB attributes to be "aliased" automatically each time new data are downloaded. In addition, any flags assigned to values in the ClimDB/HydroDB database are automatically converted to GCE toolbox flag arrays, allowing users to take full advantage of numerous flag-handling options supported by toolbox functions (e.g. selective or comprehensive removal of flagged values, display of flags above values on data plots, interactive graphical flag editing using the mouse, and automatic documentation of flagged values in derived data sets resulting from aggregation, temporal re-sampling and bin-averaging operations, with optional flagging of derived parameters based on number of flagged values in the corresponding aggregate or bin).

Mining USGS



Retrieve USGS Data

Lookup Station ID: Georgia

02317500 ALAPAHA RIVER AT STATENVILLE, GA
[illegible]
02208450 ALCOVY RIVER ABOVE COVINGTON, GA
[illegible]
02208050 ALCOVY RIVER NEAR LAWRENCEVILLE, GA
[illegible]
[illegible]
[illegible]
02390000 AMICALOLA CREEK NEAR DAWSONVILLE, GA
[illegible]

Station ID: 02226000 Data Type: Daily Days: 720

USGS Website URL: http://waterdata.usgs.gov/nwis/

Metadata Template: USGS_Doctortown

□ Save Template

☑ Close dialog after harvesting data □ Open in new window

fig 2. USGS NWIS data mining dialog
(click on image for larger version)

A very similar dialog is also available for retrieving data from any station in the USGS NWIS database (fig. 2). When the user selects a state or territory using the drop-down menu, a list of all available stations is displayed, along with short descriptions from USGS. Pressing the "Select Station" button (or double-clicking on the station entry) loads that station ID into the request form. Unlike ClimDB/HydroDB, USGS does not support querying by specific date range, so users are prompted for the number of past days of data to request. Users can request recent provisional (near-real-time, high frequency) data, recent daily data, or finalized daily data, and requests are automatically routed to the correct server and application on the USGS server. USGS limits provisional near-real time data requests to approximately 30 days and recent daily data to 720 days, so longer requests will only return the maximum allowed data records for the respective data type. In addition, finalized data are typically released mid-way through the subsequent year, so there is typically a gap of 6 to 18 months at the end of finalized data sets.

As with ClimDB/HydroDB data, users can create or modify the included metadata template to alias data columns, and any USGS-assigned Q/C flags will be converted to GCE toolbox flag arrays. Since USGS data are generally provided in English units, which are rarely acceptable for scientific work, automatic English-to-Metric unit conversions are also performed by the import filter. Both the table of unit conversions and English-Metric unit mappings can be customized by end-users using GUI dialogs, which are accessible from menus in the data set editor window.

Transforming and Integrating Data

As already mentioned, several transformations are automatically applied each time data are mined from either ClimDB/HydroDB or USGS. Some additional transformations that can be applied interactively after data are imported into the

GCE toolbox include:

- additional unit conversions and column aliasing
- deletion and re-ordering of columns
- creation of calculated data columns
- creation of additional date/time columns (e.g. YearDay from Date, Serial day from date components)
- data sub-setting or filtering based on values in any number of columns or based on mathematical expressions

Data from multiple stations can be integrated by performing pair-wise joins between data sets based on related columns (most commonly serial date or date component columns in this case), returning join columns and selected data columns from both sources in the final derived data set. In order to facilitate joining multiple data sets via a series of pair-wise joins (i.e. without first saving all data sets individually to disk), in-memory data sets can be copied to the results pane of the GCE Search Engine dialog (http://gce-lter.marsci.uga.edu/lter/research/tools/toolbox_search.htm). Using this dialog, two data sets can be selected for joining, and then the resulting derived data set will automatically be added to the list of available data sets, permitting additional join operations to be performed (in addition to other operations, such as plotting, mapping and viewing metadata contents).

A distinct benefit of using the GCE Data Toolbox for post-processing and synthesis is that all actions performed during data importing and all transformations performed by the user are automatically logged to the data set metadata, allowing the complete lineage of the data set to be viewed at any point in the process (in addition to general documentation and full attribute metadata). The user can also choose to export data in CSV or another delimited text format or copy columns to the base MATLAB workspace at any point to analyze the data using other tools or their own custom algorithms.

Application in Automated Workflows

The screen shots and descriptions in this article have emphasized interactive use of the GCE Toolbox based on GUI dialogs, which I believe would probably appeal to most potential users; however, all operations described are based on documented command-line functions that are well suited to use in automated workflow and serial batch processing scenarios. For example, data from any USGS station can be retrieved using a single call to the "fetch_usgs" function, which is described below (and displayed at the MATLAB command prompt by typing "help fetch_usgs"):

```
Fetch data from the USGS web site for the specified station and time period

syntax: [s,msg] = fetch_usgs(stationid,datatype,days,template,pn,fn)

inputs:
   stationid = string listing USGS station to collect (e.g. '022035975')
   datatype = type of data to return
      'daily' = daily summary (default)
      'realtime' = real-time data
      'archive' = reviewed, archived data
   days = number of days of data to request (default = 720 for 'daily', 31 for 'realtime', 3650
for 'archive')
   template = metadata template to use (default = 'USGS_Generic')
   pn = path to use for raw data tile (default = pwd)
   fn = filename to use for raw data (tab-delimited text;
   default ='usgs_[stationid]_[datatype]_yyyymmdd_hhmm.txt'
   baseurl = base URL for the USGS water data web site (default = 'http://waterdata.usgs.gov/nwis/')
   baseurl2 = fall-back base URL (default = 'http://nwis.waterdata.usgs.gov/usa/nwis/')

outputs:
   s = GCE Data Structure containing the requested data set
   msg = text of any error messages
...
```

Additional function calls can then be made to perform all the other transformations and integration steps described above. Command-line functions can also be combined in MATLAB scripts and other functions, and then executed on a timed basis (using MATLAB timer objects) to produce highly automated work flows that are triggered without user intervention.

Reversing the Flow: Contributing Data to ClimDB/HydroDB

In addition to retrieving data from ClimDB/HydroDB, dialogs and functions are also provided for generating harvest files for uploading new or revised data to the database. The column aliasing described above is reversed to assign accepted ClimDB/HydroDB parameter names to columns (i.e. based on entries in a user-editable column mapping table). Units are also automatically converted to standard ClimDB/HydroDB equivalents, and high-frequency data are automatically re-sampled to daily values prior to export (i.e. based on statistical aggregation using date component columns, which are auto-generated from serial date columns when necessary). Using these tools, for example, an LTER Information Manager could retrieve data from any USGS station, or import a file downloaded from the NOAA National Climate Data Center web site, and generate a valid ClimDB/HydroDB harvest file with a few mouse clicks.

Conclusion

I believe the tools described in this article, and in more detail on the GCE web site (http://gce-lter.marsci.uga.edu/lter/research/tools/data_toolbox.htm), provide a practical means for acquiring data from the ClimDB/HydroDB and USGS NWIS databases and integrating them in real time. Data from other sources, such as NOAA's NCDC web site and user-contributed data from delimited text or MATLAB files can also be imported for analysis and integration alongside long-term monitoring data from these databases.

Although these tools are freely available on the GCE web site and will run on Microsoft Windows, Linux/Unix and Mac OS/x platforms, it should be emphasized that they do require the commercial MATLAB software package in order to run, which may be a limiting factor for some potential users.

References

1. Sheldon, W.M. 2005. GCE Data Search Engine: A Client-side Application for Metadata-based Data Discovery and Integration. DataBits: an electronic newsletter for Information Managers. LTER Network.
2. Sheldon, W.M. 2002. GCE Data Toolbox for Matlab® -- Platform-independent tools for metadata-driven semantic data processing and analysis. DataBits: an electronic newsletter for Information Managers, Fall 2002 issue. Long Term Ecological Research Network, Albuquerque, NM.
3. Sheldon, W.M. 2001. A Standard for Creating Dynamic, Self-documenting Tabular Data Sets Using Matlab®. DataBits: An electronic newsletter for Information Managers. Long Term Ecological Research Network, Albuquerque, NM.
4. Henshaw, D., Sheldon, W.M. and Vanderbilt, K. 2003. Introducing the Climate and Hydrology Web Harvester System. LTER Network News, Vol. 16 No.2 Fall 2003.

File Sharing Options: Elements of a Collaborative Infrastructure

- Mason Kortz (PAL/CCE)

When working on a collaborative project, the benefits of sharing files between users quickly becomes apparent. A method has to be found that is supported by the group's collective infrastructure, that works for all members of the collaborative group, and that is (hopefully) efficient, user-friendly, and secure. This article is meant to help scope the file sharing needs of a project and find a method that meets the needs of the project, and all parties involved, without adding undue administrative or technical overhead. This article considers email, FTP, remote disks, and WebDAV as file sharing solutions.

Email, while not a file sharing technology in a strict sense, is so widely used for file transfers that it deserves some mention here. Email is a useful tool because it is ubiquitous, familiar, and in the foreground of many users' work routines. When used to transfer files, email also adds the advantage of adding context to the file, in the form of the email message body. When dealing with small, static files and small groups of recipients, email is a good choice for transferring files. However, email lacks scalability. Large attachments sent to many users can cause clutter to build up quickly. For this reason, many email services limit the size of attached files. Collaboration over email can also be difficult - with many users mass-emailing out revisions to a file, it can be hard to keep track of which, if any, is the authoritative version. Finally, emailed files don't live in a shared space. If a new user wants to start collaborating on a file, the user must first get someone who already has the file to email it out.

Another popular method of sharing files is via FTP, the file transfer protocol. Like email clients, FTP clients run the range from simple text-based clients to slick, stylish graphical interfaces. They all provide the same basic functionality - a way to move files to and from a central server. Because there is a centralized repository, there is one authoritative 'master copy' of the file being shared. A user with access to an FTP server and the proper permissions can actively find any file they are looking for; the user does not need to passively wait for someone else to send the file out. This, combined with the fact that FTP is widely supported on all platforms, makes FTP a good way to move large files between groups of users. Although storage of files is centralized with FTP, editing is done locally. In addition to being time-consuming, the upload/download mechanic means that at any time, many version of a file may exist on individual users' computers. If one user forgets to upload after making an edit, other collaborators may miss an important change.

Remote disks (also known as network drives, network volumes, sharepoints, and many other terms) are extremely useful for file sharing, especially in a collaborative environment. The exact protocols used to access remote disks vary depending on both the server and client platforms - for example, the SMB protocol is used in Windows networks, whereas AFP is more common in Macintosh networks - but they all share certain properties. Remote disks provide a centralized storage location for files, and share many of the benefits of FTP. The main difference is that remote disks, once accessed, can be treated just like local disks. This means a user can edit and interact with files in place, making the work experience smoother than with FTP, especially when working on projects consisting of multiple files. There is no upload/download or send/receive step to take up time and spread file versions around to many clients. This also means that users need to be more careful about not overwriting each others work - because edits are not done locally, any changes made to a file will overwrite the previous version. Version control and/or backup software are often used with remote disks to mitigate this problem. Another issue is that clients on many different platforms might be connecting to the same remote disk. Many applications and operating systems can leave behind special system files - for example, .DS_Store files on Macintosh OS - that have no meaning on other platforms and show up as useless clutter. This can be mitigated or prevented with the correct client settings, but this requires some extra attention from system administrators. If multiple file sharing protocols are implemented on a single server, as is often the case, system administrator overhead can increase and system security can decrease, as multiple ports must be opened.

Another protocol, WebDAV (Web-based Distributed Authoring and Versioning), bears special mention in the context of collaborative environments. WebDAV is an extension of the HTTP protocol, meaning it works through your web server, but rather than being read like a web page, a WebDAV 'site' is accessed like a file system. This allows WebDAV to use the authentication and encryption provided by the web server, and creates a secure file sharing environment without opening additional ports. As the name implies, WebDAV is designed for collaborative work. Connecting to a WebDAV share is functional very similar to connecting to a remote disk, and provides many of the same benefits. WebDAV provides a built in file-locking system that helps prevent the problem of overwrites, and resource metadata (title, authors, etc.) through XML. Future versions of the protocol will also include native version control support and more refined access control. The main downside of WebDAV is that it is a relatively young protocol, meaning that its feature set is not yet complete. Currently the permissions system provided by WebDAV is not as robust as that provided by a shared file system. Application support is currently slim and even operating system support is inconsistent. Before moving to WebDAV as the mechanism for collaboration of a project, all clients should be checked for compatibility.

So what file sharing mechanism is the best for collaboration? It all depends on the local situation. Remote disks and WebDAV provide a good collaborative environment, but require an amount of administrative overhead that, in many cases, means an over-engineered solution. For small, short-term, one-on-one collaborations, email is a fine solution. This is especially true when the collaboration is between two institutions, and getting the necessary accounts and permissions for high-level file sharing may raise administrative issues. FTP can be very useful when dissemination of files, especially large files, is more the focus than collaboration. FTP can be a good way to move files between institutions without having to create accounts, provided the one of the institutions' infrastructures supports anonymous FTP. For long-term collaboration with a large group of users, setting up a remote disk or a WebDAV share is worth the administrative overhead. Ultimately, it is best to be familiar and comfortable with several methods of file sharing, to best accommodate the infrastructure at hand, the capabilities of the collaborators, and the needs of the project.

The LTER Children's Book Series - an Opportunity for IM Outreache

- Chris Gardner (MCM)

The second installment in the LTER Children's Book Series, The Lost Seal written by McMurdo (MCM) scientist Diane McKnight and illustrated by Dorothy Emerling, is currently in press and scheduled to be released in June 2006. However, the international group of children who participated in the project can already see their contributions and learn more about the story at www.mcmlter.org/lostseal.

Several LTER sites have expressed interest in developing children's books. This provides the LTER IM community with an opportunity to stimulate public interest in science and ecology through enhanced online resources related to the book. The story of The Lost Seal describes the first ever documented encounter with a live seal in the remote McMurdo Dry Valleys of Antarctica. Following the model of My Water Comes from the Mountains, the first children's book associated with Niwot Ridge (NWT) LTER, The Lost Seal features children's original artwork along with professional illustrations by Emerling. A packet containing the story, instructions for the teacher, photos of the Dry Valleys taken by MCM scientists, and video taken at the time of the event, was sent to 2-4 grade classrooms in the United States, U.K., New Zealand, and Australia, reflecting the involvement of explorers and scientists from these countries in the Dry Valleys. The children then created original watercolor paintings and wrote comments based on the plot of the story and photographs from the Dry Valleys and surrounding regions.

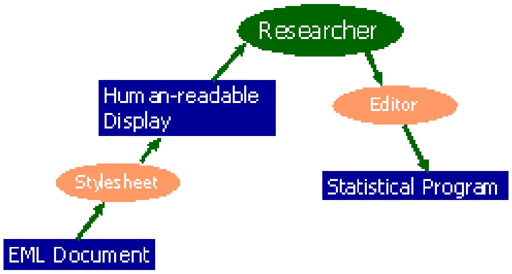
Since all of the >400 art pieces obviously couldn't be included in the book, we created a website where children can view high resolution scans of all the artwork and associated comments (www.mcmlter.org/lostseal/search.html). An undergraduate work-study student scanned the artwork and entered the comments, first names, and file paths into a spreadsheet that was subsequently loaded into a relational database. To view the art, children can either choose their school from a drop-down menu or search by their first name, either of which will take them to a JSP page that displays the query results with clickable thumbnails and comments. Another notable feature of the site is the screen shots taken from the book that are linked to videos of the actual lost seal and MCM field camps, which helps relate the story to real life events and places. Other resources on the site include the MCM photos and links to more information about Antarctica. Developing a site-level network of websites related to the children's book series would add an excellent interactive and expandable educational element to the series, as well as provide access to the children's art that was not chosen for the print version of the book.

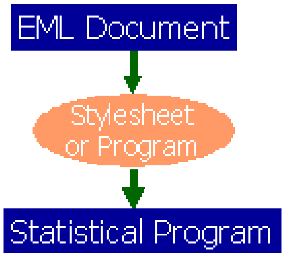
Putting EML to Work: The PTAH Project

- John Porter(VCR)

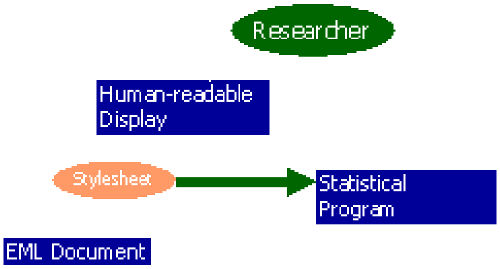
Ecological Metadata Language (EML) is a major step forward in the management and exchange of ecological metadata. By providing a consistent structure for metadata elements, it permits access to those elements by programs as well as people. Thusfar, LTER efforts have focused primarily on producing EML metadata. The "Metacat" system uses EML documents to create a cross-site data catalog, but there have been few other applications (such as Kepler) that use EML documents. Here I discuss several new applications that have been developed that exploit EML to help aid in statistical analyses.

The basic steps followed by researchers in analyzing their data have been relatively unchanged since the advent statistical programs on personal computers in the 1980s. Based on whatever (minimal) metadata is available, a researcher would use an editor to create a statistical program capable of reading the data (Fig. 1). In the case of EML documents, a stylesheet is used to translate the underlying document into a human-readable display, but otherwise the process is almost entirely the same as in the past.

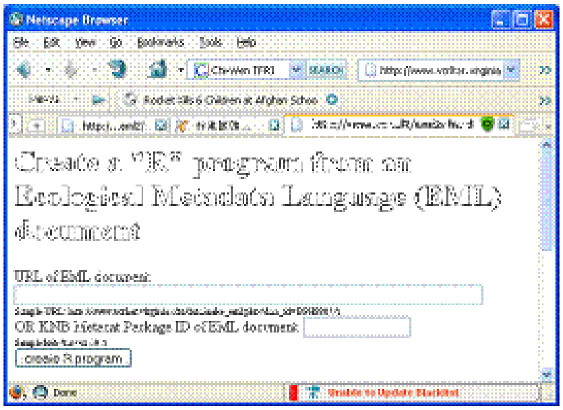




However, with the advent of EML containing attribute-level data, a new option is available - the direct generation of a statistical program that handles the routine tasks of reading the data, adding labels and missing values statements and basic statistical analyses (Fig. 2). This allows researchers to "shortcut" the laborious process of creating the basic statistical program and instead focus on the creative aspects of the analysis, by adding additional commands to the program written by the stylesheet. The EML document also plays the role of providing the researcher with the details on methodology etc. that are critical to understanding appropriate use of the data.



The PTAH (Processing Techiques for Automated Harmonization, also the Egyptian god of creation) project has as its goal the creation tools that use EML to support ecological research., including the creation of the needed stylesheets and supporting programs. The first step has been to create stylesheets that translate EML documents into statistical programs for the Statistical Analysis System (SAS), the Statistical Package for the Social Sciences (SPSS) and the "R" statistical packages. These stylesheets use the information in the dataTable, attributeList and physical modules of EML as the raw data for creating statistical programs. Through the use of the stylesheets, time to prepare a statistical program for a dataset with dozens of variables can be reduced to as little as 5 minutes. The stylesheets are available for download on LTERNET CVS web site (http://cvs.lternet.edu>). To aid researchers, who may be unfamiliar with XML stylesheets and templates and the programs for processing them, a web site at http://www.vcrlter.virginia.edu/data/eml2 provides a web interface for the


Netscape Browser
File Edit View Go Bookmarks Tools Help
SEARCH
Done
Unable to Update Blacklist
create R program

Create a "R" program from an Ecological Metadata Language (EML) document

URL of EML document:

OR KNB Metacat Package ID of EML document:

Currently 73% of LTER sites produce at least some EML documents with attribute level metadata, but overall only 31% of EML documents contain attribute-level data (D. Costa, pers. comm. 3/31/06). As more EML metadata is enhanced to achieve levels 3-6 as outlined in the "EML Best Practices" (http://cvs.lternet.edu/cgi-bin/viewcvs.cgi/emlbestpractices/emlbestpractices-1.0/emlbestpractices_oct2004.doc), the utility of this approach will gain in value. However, even now over 1,200 datasets contain the information needed to allow semi-automated processing using this approach.

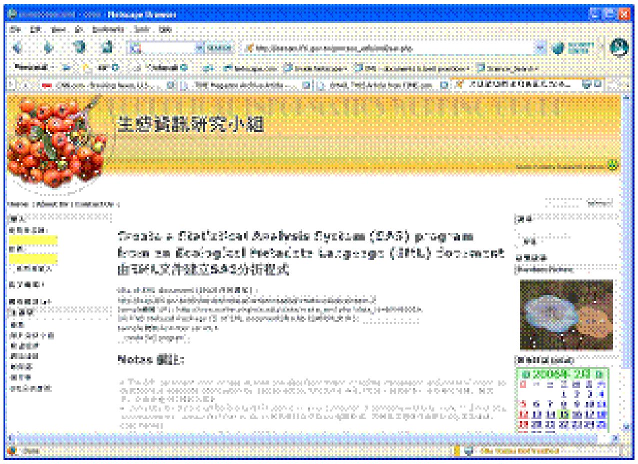
The prototype PTAH system still requires that the user edit the resulting statistical program to provide information that is not available in the EML document, such as where the data file(s) are located on the user's system. As protocols for automated, authenticated access to the raw data are developed, this step may be eliminated. Currently, the user may also need to deal with issues related to data formatting (e.g., SAS prefers missing values as character strings, even in numeric fields, but 'R' may crash if you try to read character data from a numeric field). Development of some pre-processing tools may be required to make the system more robust. Additionally, the current PTAH implementation focuses entirely on data in text files, but could be extended to use the SQL query capabilities supported in EML.

International Collaboration on EML-based Applications

- Chi-Wen Hsiao, Chien-Wen Chen and John Porter (VCR)

The Taiwan Ecological Research Network (TERN) is continuing its efforts to build Ecological Metadata Language (EML)-based systems for managing ecological data. It has engaged in several training exercises for investigators and technicians aimed at teaching them how to create EML documents. However, one of the principal questions they ask is "What do we gain by doing this?" For this reason we have focused efforts on developing some tools that use EML to create products of use to researchers. Initial efforts were aimed at data discovery, with the implementation of the METACAT system for Taiwan. Taiwan was the first LTER group worldwide to implement METACAT on a PC-based system and has also worked on resolving language and character-set issues. However, the sense is that researchers, especially those new to the concepts of ecoinformatics, want still more tools that exploit EML metadata in innovative and useful ways.

In Taiwan, the "Ecological Informatics Working Group" of TERN has developed several prototypes of tools. The first, building on the PTAH project (described elsewhere in this issue), was to develop a Chinese language interface to the programs for creating "R", SAS and SPSS programs. The interface allows researchers to use EML Metadata to create statistical program using the original PTAH engine located at the University of Virginia.



生態資訊研究小組

Create a Statistical Analysis System (SAS) program from an Ecological Metadata Language (EML) Document

由EML文件建立SAS分析程式

Notes 備註:

Subsequently, the TERN team has rewritten the PTAH engine to work on a PC-based system (the original system was UNIX-based) and extended the capabilities of the transformation tools. In particular, Chi-Wen Hsiao developed a prototype server-side system that allows researchers to access EML, upload data and then run "R" code on the server, so that researchers need not have their own copy of "R".

TERN Ecological Informatics Working Group member Chien-Wen Chen is currently on an extended visit to the VCR/LTER to collaborate with John Porter on development of sensor systems. Plans are in place for his colleagues Meei-ru Jeng and Chi-Wen Hsiao to visit later in the year to continue work on collaboration on future tool development.

Mobile GIS Technology: Trying it out

- Tom Kurkowski, Cartographer, Alaska Department of Natural Resources, Division of Forestry

Mobile GIS technology has been around for quite a number of years, but not until recently have businesses, industry, and governments begun to realize its potential. With computers getting smaller and processors becoming faster, more and more can be accomplished in the field while saving time in the office.

A mobile GIS system consists of both capable hardware as well as software specifically designed to support GIS/GPS capabilities. Mobile computers come in many varieties, but there are basically three types on the market today. Laptops are very capable machines that rival some desktop systems in system performance and have large screens that may be very useful for some field tasks. Personal digital assistants, or PDA's, are the next most popular systems due to their compactness and ability to run mobile versions of desktop software. Finally, palmtop computers are the next generation of mobile computing. These systems are a compromise between the compactness of PDAs and the speed and flexibility of laptops. They run desktop software versions eliminating the need for a separate mobile operating system. Most of these systems have ruggedized versions available ranging from a simple case to a fully sealed and shock absorbing system. Options such as additional memory, GPS, and PCI slots enabling expandability are also available. As the market realizes the advantages of mobile GIS, there will only be more choices available.

Mobile GIS software is more limited. Most software developers have realized the advantages of mobile GIS, but have delayed releasing many applications due to lack of market demand. The available applications range from simple text data collection that can synch with GIS software to software that can link up with GPS receivers and display real-time locations with satellite imagery.

When considering using mobile GIS and/or GPS technology, the specific needs of a project should be determined. Will the project only entail data collection tasks, or will querying and simple analysis also be useful? Will imagery be incorporated or useful to display during data collection? If so, what is the optimum screen size for portability and outdoor readability? Will GPS capabilities be needed?

As an example, the Alaska Department of Natural Resources, Fairbanks Division of Forestry has given me the task of developing a mobile GIS/GPS solution for wildland fire suppression as well as forest management activities. Our immediate needs are to display high resolution satellite imagery linked to a GPS in order to display real-time locational information in vehicles as well as on foot. The imagery capability limits our options because of the great amount of processing power needed to display and refresh images. We also want the ability to collect fire perimeter data by either GPS data collection or simple on screen field editing. Since the system will be used in a wildland fire situation, fully ruggedized systems are required. Additionally, the capability to expand the system to be used over a broader range of forest management activities was considered.

Currently, we are using Panasonic Toughbook laptops with touch screens that enable input through a pen, fingers, or the keyboard. Their large screen size allows satellite imagery to be clearly seen and touch screens allow simplified map navigation in the field. They are fully ruggedized and have endured field trials in snow, water, and cold and have built in GPS receivers that eliminate the need for an external receiver. The downside is the large screen size limits the outdoor readability because truly sunlight readable screen technology is limited to smaller screens, aside from what some manufactures will tell you.

In addition to laptops, we are also testing some palmtop computers for field use on foot. We currently have three oqo units that come with a five inch touch screen with pen input only as well as a miniature keyboard and mouse buttons. The five inch screen provides great portability and decent image display capabilities. The oqo runs Windows XP so any desktop software can be installed. Bluetooth technology is included, which allows wireless connections to external devices such as GPS receivers, albeit we did have problems getting the oqo to communicate with higher end GPS receivers due to oqo's proprietary bluetooth drivers. These units are not ruggedized because they were most likely designed for less demanding environments. This has led to two of our three units having major hardware malfunctions.

The mobile GIS software package we chose is ArcPad by ESRI (Environmental Systems Research Institute). This package is the most advanced available and is fully integrated with their popular desktop GIS software. This package can be installed on both Windows as well as Windows Mobile operating systems allowing one more choices in hardware selection. It is compatible with many image types including compressed MrSid and JPEG200 image formats. It allows the flexibility to simultaneously display satellite imagery with vector layers while connected to a GPS receiver which displays your current position. It allows simple queries as well as field editing capabilities. Other advanced features include georeferenced connections to digital cameras and rangefinders. ArcPad also has a companion development suit, ArcPad Application Builder, which allows customization of the out-of-the-box product. This is a great option to consider if your final goal is to develop simplified data collection programs for field personnel, except significant programming skills may be required to develop these programs.

I believe our needs for maximum portability, serious processing power, and ruggedness put our demands just past what is currently available on the market. However, most project needs of having the ability to collect data that is geographically referenced are definitely well developed in the mobile GIS market. As more and more people realize the benefits of integrating field activities with GIS, only more options will appear.

Hardware

Panasonic toughbooks: http://www.panasonic.com/business/toughbook/home.asp
oqo computers: http://www.oqo.com/
MCC Computer Company: http://www.modular-pc.com/index.htm
Itronix: http://uk.itronix-europe.com/home.asp
Xplore Technologies: http://www.xploretech.com/index.pl
Xybernaut: http://www.xybernaut.com/
DualCor: http://www.dualcor.com/

Ultramobile PC (Microsoft):http://www.microsoft.com/windowsxp/umpc/default.mspx

Software
ArcPad: http://www.esri.com/software/arcgis/about/arcpad.html
ArcPad Application Builder: http://www.esri.com/software/arcgis/arcpad-appbldr/index.html
CartoPac: http://www.spatialdatatech.com/Products/CartoPac/index.html
TerraSync: http://www.trimble.com/terrasync.shtml

Trends in Long-Term Ecological Data: a collaborative synthesis project Introduction and update

- Christine Laney & Debra Peters

Introduction: A large suite of National Science Foundation (NSF) and United States Department of Agriculture (USDA) supported long-term research sites, including the 26 LTER sites, represents a wide range of ecosystem types, from forests to grasslands and shrublands, freshwater lakes and streams, near coastal marine and estuaries as well as urban areas and systems in the arctic and Antarctica. A variety of different kinds of data have been collected from these sites through time, ranging from primarily climatic and demographic data since the 1800s to more recent quantitative assessments of plant, animal, and microbial populations and communities, hydrological and biogeochemical cycles, biodiversity, and disturbance regimes. In addition, pollen records and tree-ring data can be used to push data availability back even further. As the LTER enters its "Decade of Synthesis", the United States Forest Service (USFS) enters its "New Century of Service", the USDA Agricultural Research Service (ARS) enters the time when "The Future Grows Here", and new initiatives, such as the National Ecological Observatory Network (NEON) become operative, there is a critical need for a collection of highly accessible, up-to-date, and easy to use data sets that span the ecosystems and history of the U.S. In response to this need, the Trends project was developed as a collaborative effort between the LTER network, USDA ARS, USFS, and the National Center for Ecological Synthesis and Analysis (NCEAS).

Goals: To create a platform for synthesis by making long-term data available, and to illustrate the utility of this platform in addressing important within-site and network-level scientific questions.

Products: (1) a book to be published by Oxford University Press that focuses on trends in long-term data within and among sites, and examples that illustrate the value of long-term data in addressing important questions for a number of sites; (2) a web page containing derived long-term data and metadata that are easily accessible for synthetic analyses by a variety of users. In addition, the Trends project includes a NIS Data Module that was approved by the LTER Coordinating Committee. The data module, includes two main goals recommended by the IMEXEC: 1) to develop a dynamic webpage that generates and displays plots from derived datasets and 2) to determine the feasibility of using EML in the development of NIS modules for solving general synthesis problems. The LTER Network Office (LNO) will assist in meeting these data module goals.

Aims:

1. Short-term (Sept. 2006): The book should be 80-90% complete, and a web page should be available to interested users in time for the All Scientists Meeting. This initial web page will contain static derived datasets, and simple search/query, download, and graphing capabilities.

2. Long-term (2007-): We will work with LNO to develop a Trends web page containing dynamic datasets that can be harvested from each site periodically, converted to derived form, searched, and available for download, graphing, and statistical analyses. We would also like to link with the education community through web sites such as Teaching Issues and Experiments in Ecology (TIEE, http://tiee.ecoed.net/).

Types of data being collected, and current status: Trends data and graphs will be organized into four themes that correspond to the LTER Planning Process: climate and physical variability, including disturbances; human population and economy; biogeochemistry; and biotic structure. Within each theme, the Trends Editorial Committee identified key variables useful for cross-site and network-level analyses. These variables will be plotted on monthly or annual time scales. In addition, graphs from each site will be used to illustrate the importance of long-term data in addressing ecological problems. We expect that most of these figures will already have been published. Initial cross-site synthetic analyses related to global climate change and human impacts on the environment will be investigated as part of the book.

Data sources:

- The National Climate Data Center's website, http://www.ncdc.noaa.gov/oa/climate/onlineprod/drought/xmgrg3.html, has been an invaluable source of Palmer drought Palmer drought severity index data for all of the continental sites. Similar types of data are needed for sites in Alaska and outside of the U.S..

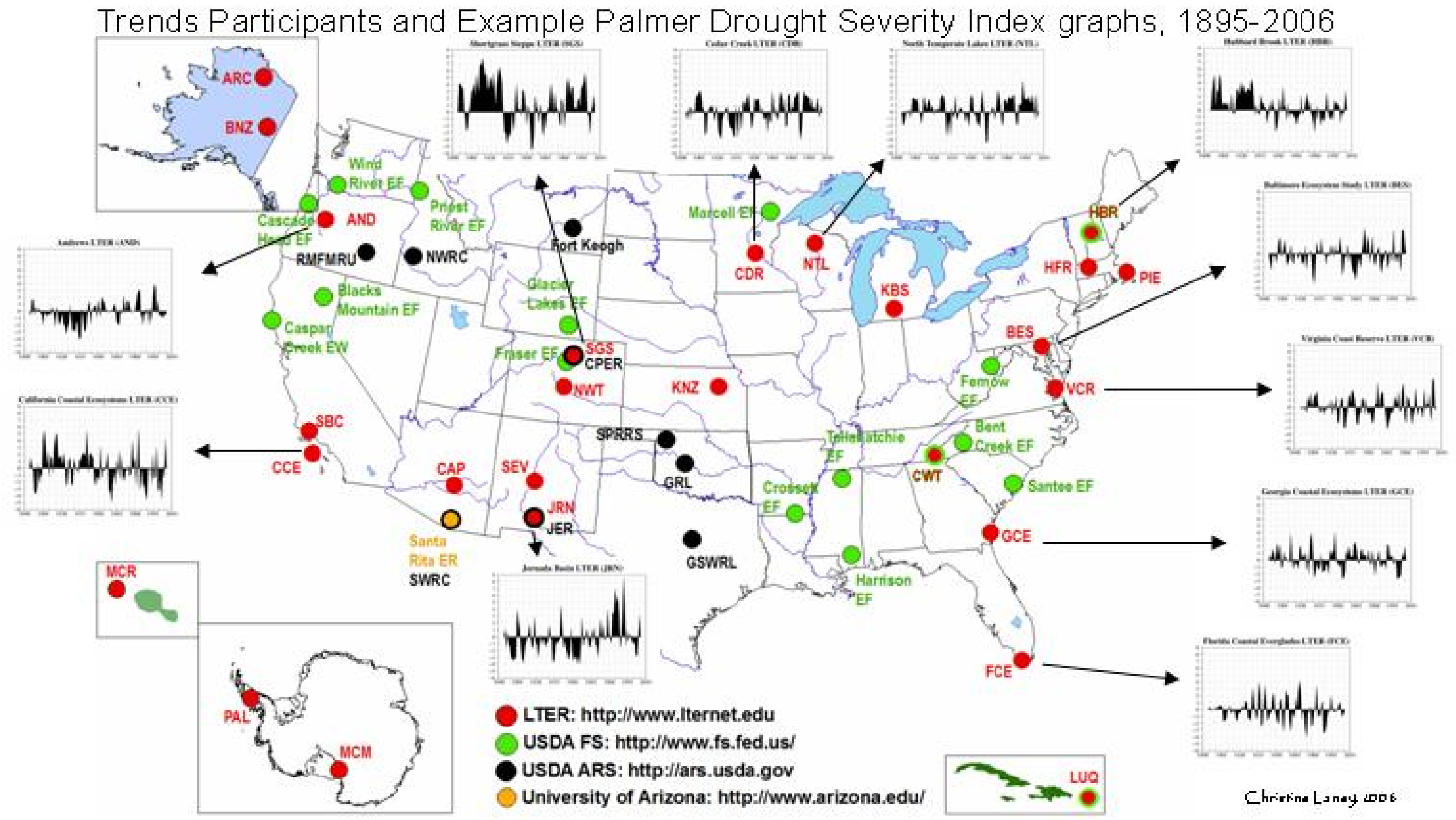
- ClimDB/HydroDB: We have extensively utilized this website to locate and download derived (monthly or annual) precipitation, air temperature, water temperature, solar radiation, wind speed, streamflow, and other available data for multiple LTER and USFS sites. The ability to obtain derived data from LTER and USFS sites in the same format, with the same attribute labels, has been invaluable. This allows us to write modular plotting scripts (we currently are using R) that can be used to quickly and efficiently plot identical graphs between sites.
- KNB and LTER Metadata catalog servers: Some data were downloaded via these servers. However, we found it difficult to search for datasets that contained 10 or more years of data using this method. We found that we also had to look at each individual site's website and/or ask the PIs or IMs directly to help us figure out which of their datasets fit the time requirements, were pertinent to the project, and contained unrestricted data.
- Site websites: LTER websites provide a great deal of information, and more is available with each passing month. EML provides a great standard format for reading and interpreting the metadata, and EML that includes attribute-level information is the most helpful. Non-EML data is often very useful as well, but as each site uses a different layout, it will require more work to utilize these metadata files.
- Information managers and principal investigators have often directly submitted data that was either not available on a website or in a metacat server, or was not up-to-date online. They have also submitted graphs, previously published figures, photos, text descriptions of their sites, and other useful information. The data has been submitted in various formats, ranging from simple fixed-width format to multi-page excel spreadsheets or even database files.

Current Status:

- *Participants:* All 26 LTER sites, 9 USDA ARS sites, 14 USFS sites, and a site administered by the University of Arizona are currently participating in the Trends project. See the figure below for coverage.
- *Climate, Biogeochemistry, and Biotic Data:* We have well over 300 datasets currently in hand, and metadata for about half of these datasets available in EML. We are currently generating some derived datasets and graphs, using R, an open source statistical program. These scripts will be stored for informational purposes, and may also be used in the future to help generate new datasets from newly harvested raw data. We are also producing EML files to document the derived data and graphs were generated. Recently, we reviewed all of the data in hand or available online to see which of the data for all of the variables are represented by each site. We have asked each site to review these data and make changes or additions to their submissions.
- *Population and Economy:* Some of the long term data included in the Trends Project will focus on Human Population and Economy of each site. Using lists of counties representing the region associated with each LTER site, long term changes in population, economic structure (employment by type), and land use will be quantified. Counties were the chosen unit of analysis because they are the only long term geography that is available for these data. Data from 1790 to 2000 will be organized from the U.S. Census, with Chris Boone (ASU: BES & CAP) working with the data from 1790 to 1970 and Ted Gragson (UGA: CWT) using the data from 1970 to 2000. Trends in land use provided by each site will also be included. A prototype for this section was developed by Boone and Morgan Grove (USFS: BES) using data from BES.
- *Database:* We created a database to hold basic information about each dataset, along with the locations of each dataset's data files, metadata, graphs, etc. This database will eventually be used by the LNO-developed module to allow users to access the data and metadata.

Future events:

- *Editorial committee meeting:* The editorial committee will work on final formatting and organization of the book and initial web page in July.
- *All Scientists Meeting:* Christine Laney and Ken Ramsey will host a working group session in the afternoon (1300 - 1600) of the ancillary day, Sept. 20. They will focus on IM issues associated with synthesis projects like Trends. Deb Peters and Christine Laney are hosting an evening session on Friday, Sept 22, to present Trends to the broader LTER community. At this point, additional data and graphs can still be added to the book and web page from the community


Trends Participants and Example Palmer Drought Severity Index graphs, 1895-2006
Shortgrass Steppe LTER (SGS)
Cedar Creek LTER (CDR)
North Temperate Lakes LTER (NTL)
Hubbard Brook LTER (HBR)
Baltimore Ecosystem Study LTER (BES)
Virginia Coast Reserve LTER (VCR)
Georgia Coastal Ecosystems LTER (GCE)
Florida Coastal Everglades LTER (FCE)
Jornada Basin LTER (JRN)
Andrews LTER (AND)
California Coastal Ecosystems LTER (CCE)
ARC
BNZ
Wind River EF
Cascade Head EF
AND
Priest River EF
RMFMRU
NWRC
Fort Keogh
Marcell EF
CDR
NTL
HBR
HFR
PIE
KBS
BES
Blacks Mountain EF
Caspar Creek EW
Glacier Lakes EF
Fraser EF
SGS
CPER
NWT
KNZ
Fernow EF
VCR
SBC
CCE
SPRRS
GRL
Tallahatchie EF
Bent Creek EF
CWT
Santee EF
CAP
SEV
JRN
JER
Crossett EF
Santa Rita ER
SWRC
GSWRL
GCE
Harrison EF
MCR
PAL
MCM
FCE
LUQ
● LTER: http://www.lternet.edu
● USDA FS: http://www.fs.fed.us/
● USDA ARS: http://ars.usda.gov
● University of Arizona: http://www.arizona.edu/
Christina Laney 2006

Multiple Approaches to Semantic Issues: Vocabularies, Dictionaries and Ontologies

- Karen Baker (PAL/CCE), Deana Pennington (LNO), John Porter (VCR)

A need for collaborative bridging between some of the multiple LTER approaches to semantic work was recognized at the LTER Information Management Meeting in 2005. Our intent in this issue is to summarize ongoing activities and to prompt individual insight and dialogue between LTER Information Management Meeting (2005) working groups working on semantic issues as well as within the Information Management and LTER communities in general. The series of three articles that follows represents interrelated efforts that contribute to information infrastructure building.

1. Improving Data Queries through use of a Controlled Vocabulary
2. Creating Information Infrastructure through Community Dictionary Processes
3. Navigating Semantic Approaches: From Keywords to Ontologies

These articles represent differing temporal scale approaches to community infrastructure development. The Controlled Vocabulary Working Group (article 1) identifies immediate tasks for a series of sub working groups; the Dictionary Working Group (article 2) focuses on a short-term product but a long-term collaborative process; and the Science Environment for Ecological Knowledge (SEEK) Knowledge Representation group (article 3) highlights the long-term nature of their work: "The challenge of leveraging these different approaches is one of navigating different temporal scales of development such that we can clearly envision future linkages and work independently towards a collective goal." This opens up the notion of a work arena where teams focus on a variety of semantic approaches as part of the community knowledge-making that encompasses data collection to knowledge representation - where the aim is not isolated information announcements but rather ongoing loosely coordinated exchanges.

Such integrative efforts provide an opportunity to step back to ask "What does an information system that stretches from field site to community and national centers look like"? The LTER Information Management Committee benefits from the shared experience of the long-term implementation effort with the Ecological Metadata Standard. From this experience, we may learn the value of discussing 'What is a standard and what are its ramifications and timeframes?' in addition to addressing the action-oriented 'How do we implement a standard?'. Analogously we benefit now from asking jointly 'What are the elements of an LTER information infrastructure?' and 'What are the multiple dimensions and timeframes of LTER semantic efforts'? Such inquiries require community discussion in order to provide input on both framing and reframing questions and on creating processes to begin formulating their solutions. Semantic work challenges call for development of an assortment of strategies and collaborative mechanisms - all as part of a coordinated information infrastructure stretching from the immediate to the long-term.

Improving Data Queries through use of a Controlled Vocabulary

- John Porter (Virginia Coast Reserve LTER)

Currently the keywords used to characterize datasets at most LTER sites are uncontrolled, meaning that they are selected entirely by the data creator. One of the challenges facing LTER and external researchers in discovering data from LTER sites is inconsistent application of keywords. A researcher interested in carbon dioxide measurements must search on both "Carbon Dioxide and "CO2." Moreover, a search on "Gases" would not find either of them.

The existing set of words and multi-word terms is highly diverse. For example, in the EML documents comprising the LTER Data Catalog, over half (1,616 of 3,206) the keyterms are used in only a single instance. Only 104 of the terms are used at 5 or more different LTER sites. The situation is similar for other lists of words:

Source	Number of Terms	Number used at 5 or more sites	Frequently used words
EML Keywords	3,206	104	LTER (1002), Temperature (701)
EML Titles	2,825	213	And (768), Data (482), LTER (378)
EML Attributes	6,318	436	The (4,207), Data(1,621), Carbon(328)
DTOC Keywords	2,774	103	ARC (1645), Temperature (732))
Bibliography Titles	13,538	1,855	Of (12,611), Forest (2,050)

To help improve this situation, a working group at the 2005 LTER Information Managers' meeting met to develop a plan for improving the searchability of LTER data. The plan revolves around identifying existing controlled vocabularies, thesauri and ontologies that could be exploited to help provide a richer content for searching LTER data. The working group came up with a three part plan:

1. **Information gathering:** Accumulate and analyze lists of words and terms used by LTER researchers. Combine these lists and identify a set of "important" terms that can be used to test the richness of existing resources. Words from existing site-specific controlled vocabularies will also be gathered for use in the testing phase. This list-gathering phase of the plan is largely complete thanks to the efforts of Duane Costa (EML lists), James Brunt (Bibliographic list) and John Porter (Data Table-of-Contents/DTOC list), and the lists are posted on the http://cvs.lternet.edu web site. These individual lists were then combined to produce a consolidated listing of 21,153 words or terms along with:
   - Number of lists on which it appeared (range 1-5)
   - Number of sites and uses from each list (EML Title, Keyword and Attribute; Bibliography; and DTOC)
   - Max and Min number of sites using within a list (0-24)
   - Max and Min number of uses within a list (0-12,611)
   - Is it a multi-word term?

   This consolidated listing is now available via the LTER Metacat as package knb-lter-vcr.147.1, and is free for use by other ecoinformatics groups interested in analyzing LTER content. During a videoconferencing session in April, a sub-working group chaired by John Walsh and Barbara Benson was charged with the development of one or more (shorter!) lists where words or terms are rated in terms of their "importance."

2. **Testing:** The goal of this step is to use the list(s) of "important" words and words from site-specific controlled vocabularies to test the utility existing lexigraphic resources such as controlled vocabularies, thesauri and ontologies. These resources will be rated based on the number of "important" words that are found in a given resource along with measures of how "rich" the information that resource returns, such as number of more general terms, more specific terms or related terms. During the April videoconference, a sub-working group chaired by Inigo San Gil was charged with deciding: a) what should the content of a "report" from a test session include?, b) which resources should be evaluated? and 3) how should the testing be conducted? When the work of this subgroup is completed, we should have the information needed to make decisions about which lexigraphic resources are likely to be most useful.

3. **Development:** Once the lexigraphic resources (existing controlled vocabularies, thesauri and ontologies) have been evaluated, one or more will be selected for utilization by LTER. This may involve negotiating formal Memoranda of

Understanding (MOU's) with the resource creators or working with them to enrich their content to support LTER searches. Additionally, tools that use these richer information resources need to be developed, so that users searching for data will have access to improved search and browse tools. During the April videoconference, the LTER Network Office agreed to take the lead on developing prototype applications that are capable of using a wide array of lexigraphic resources. However, initially they will be tested using a smaller subset of resources while the information gathering and testing phases are completed.

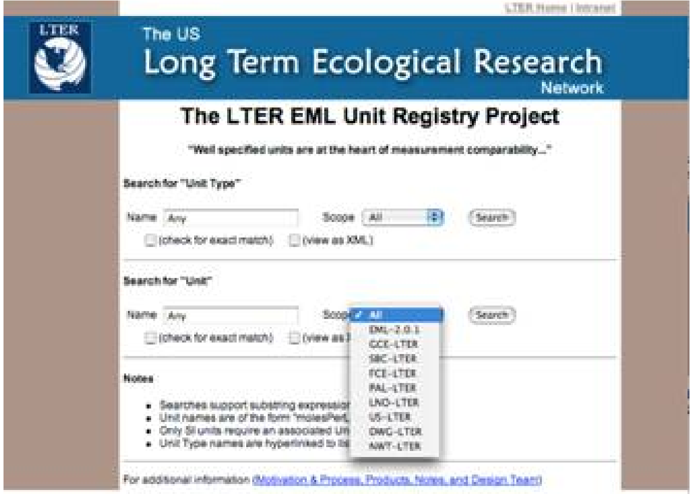
The sub-working groups aimed at evaluation of lists (identifying "important" terms), testing against existing lexographic resources and development will be working over the next several months on their respective tasks. Information managers and others interested in participating in the sub-working groups should contact the sub-group leaders.

Creating Information Infrastructure through Community Dictionary Processes

-Karen Baker*, Lynn Yarmey*, Shaun Haber*, Florence Millerand*, Mark Servilla** *PAL/CCE; **LNO

A Dictionary Working Group meeting and prototype demonstration were held during the LTER Information Manager 2005 Meeting. The group formed in order to consider and broaden the work of a Unit Registry Design Team that collaborated periodically in the months preceding the LTER IM Meeting. This effort included participants spanning five sites and the network office (KBaker, MServilla, LYarmey, LPowell, MO'Brien, WSheldon, SHaber, FMillerand, ISan Gil). From the demonstration and working group discussion, several information managers explicitly expressed interest in joining the working group (TAckerman, HGarrett, JWalsh). The group's aims may be summarized as: preparing a prototype, developing a dictionary process, and creating a community process. Having already prepared a prototype unit registry, the group's next steps include assessing the prototype and developing federating mechanisms between local and community dictionaries for adding new units and unit types. Additionally, the development of an attribute (or parameter) dictionary is of parallel interest to address the organizational needs at the local level for integrating dictionaries. Capturing and embedding the interdependence

A dictionary is a list of words with information about their definitions and characteristics. Having a unit registry associated with an LTER unit dictionary is a design feature that transforms the dictionary from a passive static list to an interactive community tool - a living dictionary. An important aspect of the unit dictionary is the concept of 'scope'. The scope represents a strategy to define the working acceptance of a specific unit within the community. For example, when a site submits a term, it is flagged as "site-level" (e.g. PAL-LTER). Site-level implies that the unit is accepted only at the site for use in their dictionary process. Units that are under Network review by the dictionary working group are designated as 'DWG', approved at the Network level as 'US-LTER', and at the community level as 'EML'. The initial unit dictionary included all units defined within the EML 2.0.1 specification, and thus have an EML designation. Such a plan is technically straightforward though it would take extensive community and organizational work to enact. A view of the prototype is included in Figure 1.



The LTER EML Unit Registry Project

"Well specified units are at the heart of measurement comparability..."

Search for "Unit Type"

Name: Any Scope: All (Search)

□ (check for exact match) □ (view as XML)

Search for "Unit"

Name: Any Scope: ✓ All (Search)

□ (check for exact match) □ (view as

Notes

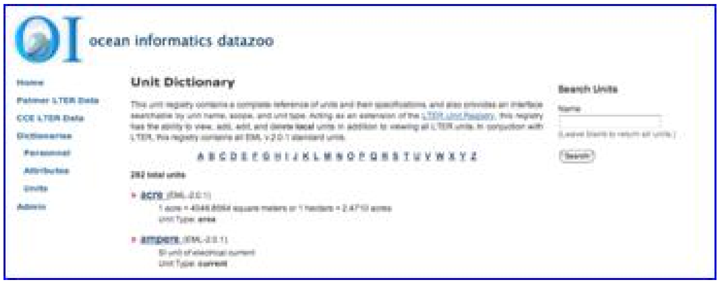
- Searches support substring expressio
- Unit names are of the form "molesPerL
- Only SI units require an associated Un
- Unit Type names are hyperlinked to lis

For additional information (Motivation & Process, Products, Notes, and Design Team)

The prototype is described in the meeting report (http://gce-lter.marsci.uga.edu/lter_im/2005/index.htm) and available online (http://fire.lternet.edu/customUnit) as is the unit registry code (http://cvs.lternet.edu/cgi-bin/viewcvs.cgi/NIS/projects/customUnit/).

The initial focus of the LTER Dictionary Working Group was to use a participatory design methodology for considering the dictionary process (see Living Dictionary, Databits Spring 2005). Our first case study yielded the unit registry prototype. One lesson learned from this work is the importance of recognizing the addition of new terms to a dictionary as a community negotiation and learning process. This means that issues like duplicate entries are not bugs, but they are markers of agreements-to-come. At minimum they serve as dialogue prompts, and at best they produce knowledge making moments where local everyday assumptions are considered and differences revealed as options to discuss.

Subsequent to the summer LTER IM Meeting, the unit registry code was checked out from the CVS Repository to a local site (PAL) in order to test its portability and to investigate integration with a multi-project site attribute dictionary. This migration, if viewed as a robustness indicator or prototype benchmark, was successful in practice. Local efforts were able to leverage without any major difficulty from the previous unit registry work. Additionally, a web interface has been developed to enable input as well as administration of unit and attribute dictionaries (see Figure 2). Further working group discussion and resource arrangements will decide the trajectory of this work.



Unit Dictionary

This unit registry contains a complete reference of units and their specifications, and also provides an interface searchable by unit name, scope, and unit type. Acting as an extension of the LTER Unit Registry, this registry has the ability to view, add, edit, and delete **local** units in addition to viewing all LTER units. In conjunction with LTER, this registry contains all EML v.2.0.1 standard units.

282 total units

- acre (EML-2.0.1)
  1 acre = 4046.8564 square meters or 1 hectare = 2.4710 acres
  Unit Type: **area**
- ampere (EML-2.0.1)
  SI unit of electrical current
  Unit Type: **current**

(Leave blank to return all units.)

Data dictionaries provide a mechanism to gather and preserve information about field observations as well as to inform both data collectors and data users. Unit and attribute dictionaries represent an organizational strategy and are one element of an information infrastructure. As we gain experience with the scope of our local and community data as well as with information classification, we begin to build an understanding of data typologies, units, and attributes. The process of creating dictionaries establishes a unique setting for dialogue between information system requirements, information managers, and earth scientists.

A dictionary can create a bidirectional forum - one of both elicitation as well as prescription. It serves as a mechanism prompting self-organization; with an explicit organization, it exerts control. What's in the dictionary informs, yet is subject to discussion and update when appropriate processes are in place. So after all, is a dictionary just a controlled vocabulary list? Or is it a moderated forum informed by community needs? Dictionaries are an infrastructure element that may be enhanced by technical structure, organizational flexibility, and community use.

Communication among the three LTER Working Groups - controlled vocabulary, data dictionary and knowledge representation - was facilitated by a visit of Deana Pennington to UCSD/SIO on 16 March this year as part of an ongoing series of Ocean Informatics Exchange events. Ocean Informatics is a conceptual framework for marine science information management efforts including the Palmer LTER, the California Current Ecosystem LTER, the California Cooperative Fisheries Investigations as well as some elements of the Southern California Ocean Observing System (SCCOOS). A hands-on workshop with Ocean Informatics participants created an active forum for data-to-knowledge mental maps; a cyberinfrastructure presentation to the LTER CCE community provided a broad context for local efforts. The visit included a conference call with John Porter as lead of the Controlled Vocabulary Working Group and discussions about LTER related social informatics and articulation work supported by the ongoing Human Social Dynamics Comparative Interoperability Project. The meeting was a continuing recognition of the need for new forums and formats that cross projects, institutions, and traditional task structures. In overload with existing requirements for collecting, managing, and publishing data, there is typically a lack of organizational support and of resources to dedicate to creating new structures that function as infrastructure building information exchanges.

With data collection and research ongoing, there is a question of where to start in organizing efforts for wider data sharing activities. A dictionary is one place to start. It creates a language held in common at a site. Merging a site dictionary with a community dictionary may be viewed as problematic or, alternatively, as creating an opportunity, a venue for discussion and negotiation. The dictionary becomes a mechanism to make visible the process of information sharing and of community arrangement making. The dictionary initiative emerged in the midst of EML implementation efforts, when information managers were looking for tailored pragmatic solutions to facilitate the transition of local arrangements to support EML. From the Dictionary Working Group development process emerged something more than an additional unit registry to augment the EML unit dictionary. The Working Group served as a coordination tool between the sites themselves and between the sites and the LNO, initiating a community process upon which further information infrastructure efforts may build.

Dictionaries are one of a suite of semantic tools for developing local and federated information infrastructure. In the semantically rich and chaotic realm of observational research, data dictionaries serve as a point of engagement for participants in preparing for data sharing. They provide a place to start for data collectors to engage with community expectations that are semantically demanding and to align with complex information system requirements.

Navigating Semantic Approaches: From Keywords to Ontologies

- Deana Pennington - LTER Network Office and SEEK Project

Ontologies! Controlled vocabularies! Data dictionaries! These and a multitude of other terms are coming into widespread use as we grapple with semantic methods for clarifying the meaning of words used to describe computational resources. How ironic that a field whose goal is semantic clarification is itself littered with unclear terms! In actuality, the terms have precise meanings, and well-understood implications. The goal of this article is to provide a conceptual framework for understanding these different semantic approaches, introduce the approach being used within the Science Environment for Ecological Knowledge (SEEK) project, and suggest some opportunities for leveraging different approaches ongoing within LTER.

Knowledge representation (KR) is a very broad field. In its most general sense, it is simply methods for external representation of the things that we know internally. There are things that exist (physical and non-physical/abstract); we represent them in different ways in order to be able to talk about them. Natural language is a form of knowledge representation - we assign words to represent things that exist in the world. Mathematics, and physical or computational models, are ways to represent knowledge. The same thing can be represented many different ways, and the choice of representation will affect the ways in which one can talk about and/or reason about things. All representations are necessarily imprecise and inaccurate because the only completely precise and accurate representation of a thing is the thing itself. The best representation depends on the objective.

Here, we are primarily interested in a technology view of KR, where the goal is to provide automated reasoning regarding semantic compatibility of resources (data, computational models, etc.). For example, using John Porter's example (this issue), we would like the system to be able to determine through automated reasoning that there are semantic relationships between the terms "CO2", "carbon dioxide" and "gases," and to be able to perform different tasks based on the degree to which those relationships are specified. Choices in representation methods are primarily choices about to what extent and how the relationships are specified, and there is a trade off between the degree of expressiveness and automatic reasoning capability. For example, natural language is very expressive, but does not lend itself to any kind of automated reasoning

Common methods used to achieve semantic clarification are shown in Table 1, along with their characteristics. First, we recognize that a thing (physical or abstract) must be represented by some kind of symbol. The set of like things form a concept which we can define and may represent with multiple symbols. For instance, all carbon dioxide molecules through time constitute a set of like things (the molecules themselves) that we perceive as a single concept that we can explicitly define and that we represent with multiple terms (CO2, carbon dioxide). Synonyms are terms that represent the same concept. Any additional structure that we impose on a group of terms comes from defining other relationships between terms. Classification is the process of organizing a group of concepts into a subsumption hierarchy where the relationship between two concepts is in the form of broader and narrower terms (e.g. CO2 "isa" gas, where "isa" is the conventional way of referring to this relationship in KR). Properties are the defining characteristics of each concept, which requires linking two concepts with a "has" relationship (e.g. CO2 hasProperty odorless, where "odorless" is its own defined concept). Classification occurs through property assignment, therefore these two approaches go together even though property assignment may be tacit rather than explicit. For instance, the concept "gas" has certain properties by which you determine which things belong in that concept. The concept CO2 is a subset of the concept gas, and as such all things that are represented by the concept CO2 inherit the properties of the concept gas, but those properties are further restricted by additional properties of CO2 that not all gases possess. Hence, classification is a process of organizing concepts by properties, whether or not those properties are made explicit. Lastly, other kinds of relationships between concepts can be stated. However, logical reasoning is based on the mathematics of set theory, therefore automated reasoning engines typically operate on "isa" and "has" relationships. This seems straightforward but in practice, appropriately structuring concepts such that logical reasoning can be optimized is a challenge. There are common pitfalls in organizing subsumption hierarchies that can negatively affect reasoning capability (perhaps the subject of a future article).

Table 1. Methods of semantic clarification and attributes of each. Parentheses () indicate an attribute is sometimes incorporated into the given method, but not required.

	Definition	**Synonyms**	**Classification (isa)**	**Properties (has)**	**OtherRelations**
Keywords					
Dictionary	X				
Controlled vocabulary	(X)	X			
Thesaurus	X	X			
Taxonomy	(X)	X	X	(X)	X
Ontology	X	X	X	X	X

A key issue in the choice of semantic method is the level of automatic functionality provided (Table 2). Any added semantic information, at any level, will enable better resource discovery. Simply assigning keywords (often called annotating) to resources is much more efficient than searching through the entire text of the resource. Providing a defined list of keywords (data dictionary) clarifies what terms may be searched and what those terms explicitly mean. A controlled vocabulary limits the terms that may be used. A common usage of controlled vocabularies is to avoid the use of synonyms such that a search on a single term should yield all relevant resources (e.g. either CO2 or carbon dioxide would be provided as a keyword, but not both). A thesaurus allows synonyms and specifies the link between them. In addition, a thesaurus usually links words that are related but not synonyms, and sometimes links antonyms. Searching for resources via a thesaurus would return those annotated to the term you searched for and those annotated to any related terms. For instance, a search on carbon dioxide would return datasets annotated with the keyword CO2, even if you didn't specifically request that term. A taxonomy adds a classification hierarchy, therefore a search can include narrower terms. For instance, a search on the term "gases" including narrower terms would return resources annotated to the term CO2 even though CO2 was not specifically requested. An ontology provides similar functionality, but additionally one could search by properties, for instance, a search on "gases" that have the property "odorless" would return a resource annotated as CO2.

Table 2. Methods of semantic clarification and attributes of each. Parentheses () indicate an attribute is sometimes incorporated into the given method, but not required.

	Keywords	**Dictionary**	**Controlled vocabulary**	**Thesaurus**	**Taxonomy**	**Ontology**
Discovery	X	X	X	X	X	X
Intergration				X	X	X
Working Analysis						X

In addition to resource discovery, semantic clarification can aid in automating data integration. For instance, if two datasets have the same format and structure, but one dataset is annotated CO2 and another carbon dioxide, then a system linked to a thesaurus, taxonomy, or ontology could automatically join the two datasets into one. Or, if one would like an integrated dataset of all information on dissolved gases in a given lake, a search on "gases" linked to a taxonomy or ontology would return a dataset annotated with the concept CO2 that could then be automatically integrated with other datasets, depending on other system criteria.

Ontologies combined with automated reasoning can enable a broad array of more sophisticated functionality. Rather than attempt to describe that functionality in general terms, it is likely to be more useful to describe one specific example that illustrates the level of functionality that could be obtained. In SEEK, we have a biomass example that we have used repeatedly to ground our understanding of the practicalities of ontology usage (Figure 1). Given two datasets, one that contains information about plant species, cover area, and height, and a second that contains information about plant species and biomass, we "know" that there is a relationship between these two datasets (at least, those of us who work with plant biomass). The reason we know that, is we have a conceptual model in our head that can easily make the inference that we can calculate plant volume from area and height information, and that plant volume is related in context-specific ways to biomass. If we know that volume equals area times height, and we know the function to transform plant volume to biomass, then we can integrate these two datasets manually. However, if we have 1000 datasets of each and no devoted graduate research assistant on which to confer this task, this is not a desirable approach.

The general approach to automate this task duplicates our own reasoning approach. We need to formally encode a conceptual model in a language that the computer can understand (an ontology), specify how each concept that is in our datasets fits into that conceptual model, and find the algorithms on the system (tools) that can do the transformation between those concepts (Figure 1). This is equivalent to following the path from AREA and HGT in the first dataset, to corresponding concepts in the ontology, following those to the Calculate Volume tool, using the output from that as input to the Calculate Biomass tool, then integrating the result with the second dataset (Figure 1).

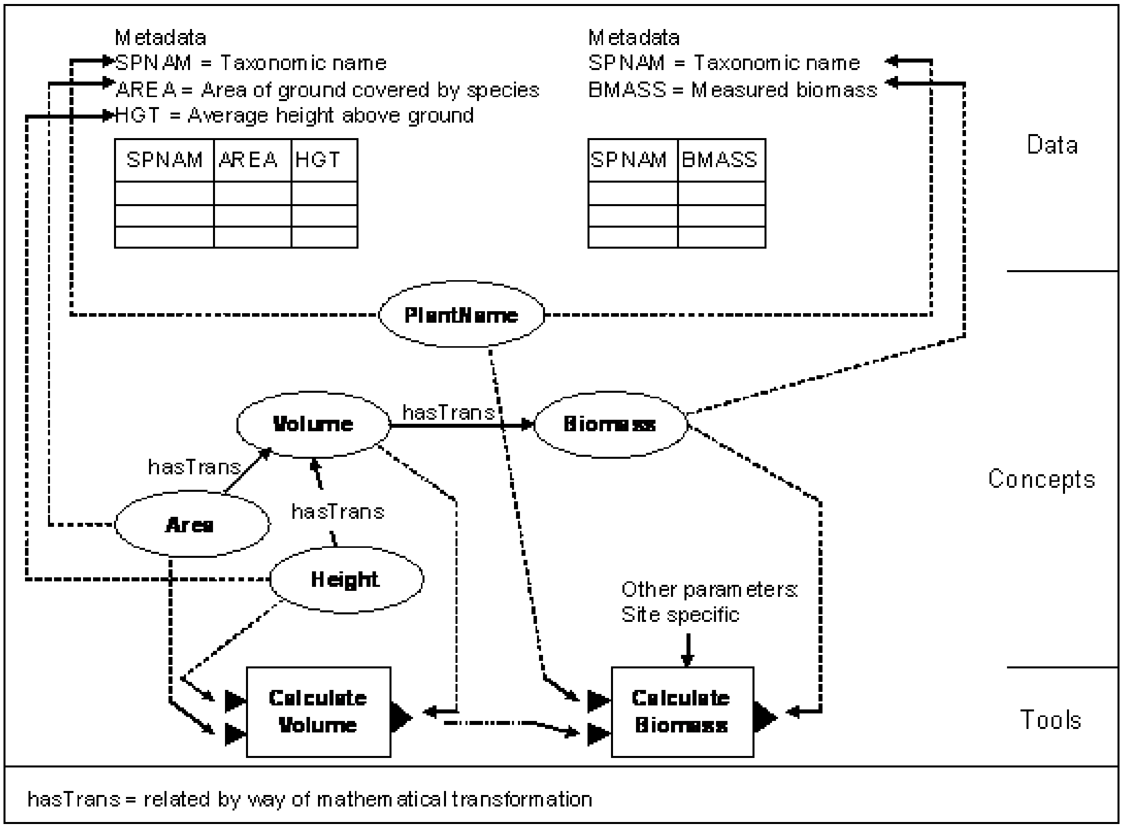


Figure 1. Illustration of the use of ontologies to automatically integrate annotated datasets by way of annotated tools, the use of which can be logically inferred by a reasoning engine.

This fairly straightforward task explodes into complex details when implemented on the system. Logically, the system has to make two connections before figuring out the path. It must recognize that it needs plant volume to get to biomass, and that it can get plant volume from area and height of plant species. It must infer a couple of dozen automatic steps to accomplish that, including finding each concept occurring in the datasets in the ontology then figuring out how they are related, what tools are related to the same concepts and which one(s) will make the correct transformation. What if there are another dozen columns not shown here? The system must locate all of those and determine any and all relationships between all column concepts even if they aren't relevant to the immediate question (although the system can allow the user to specify which columns are needed, often all columns are integrated in order to keep the extra information).

The five explicit concepts used by the resources in question (data and tools) can be represented by different terms and organized in different ways. For instance, volume, area, and height are clear terms with clear mathematical relationships, but should we have a separate concept for plant volume and relate that to biomass, or do we force the system to figure out that it is plant volume by recognizing (somehow) that the area and height columns are related to the species name column, which contains only plant species? What is the right trade off between ontology complexity versus system complexity? Do we have a separate section of the ontology specific for plant measurements and link those to generic measurement terms? What are the implications of our choices on other applications that we have not yet thought about? This is a single example of functionality that could be provided by an ontology - there are many others that use the ontology in different ways, all of which bring a set of complexities into the game.

To address these complexities, SEEK established a KR team made of a number of people who participate at different times in different ways, but whose regular participants have been Mark Schildhauer (NCEAS), Josh Madin (NCEAS), Shawn Bowers (UC Davis), Ferdinando Villa (U Vermont), Sergei Krivov (U Vermont) and myself. Determining an appropriate generic framework from which to begin population of an ecologic ontology has occupied many grueling hours of discussion by the team. Progress is slow and includes hours of abstract, philosophical discussion. For example, we recently spent two days discussing what, exactly, is an observation, how does that relate to columns in a dataset, and what is the philosophically correct way to represent this in an observation ontology that will allow for re-use of the data in ways that were not necessarily intended by the data collector? Is the entire row a single related observation or is each column an independent observation, and what are the implications of that decision for automated reasoning? What about columns that hang together, like columns designating blocks and replicates of a field experiment? Are those spatial concepts, experimental concepts, or some kind of hybrid, and how would you represent that in a generic ontology?

There are few analogous ontologies that have been developed and applied in other disciplines. Prior experience in medical and biologic domains has been limited mostly to integration of synonyms and hierarchical concepts. For instance, the gene ontology that has been highly effective is limited to expression of a relatively few categories of concepts using very few kinds of relationships. Ecology, in contrast, is a science about a multitude of relationships between many different kinds of concepts, and any ontology that will be useful for the applications we have will necessarily be complex.

Ontology development in ecology is going to be a long, complex process. In the meantime, simpler, quicker methods such as data dictionaries and controlled vocabularies have an important role to play. They can greatly enhance our ability to discover and make sense of relevant information. They can inform ontology development by providing lists of relevant terms with which to populate the ontology once a framework is in place. They can provide paths by which resources could be (semi-)automatically annotated to ontologies. Conversely, once ontologies are in place, they can be used to inform simpler methods. For instance, the decision to add a term to a controlled vocabulary could be informed by displaying terms that might be related. The challenge of leveraging these different approaches is one of navigating the different temporal scales of development such that we can clearly envision future linkages and work independently towards a collective goal.
Many opportunities for collaboration exist, but many barriers to working across scale, institution, organization, and culture exist as well. The challenge of the human dimension may well be more difficult to overcome than any of the technical challenges

Editorials

****Editor note - There are many great database programs and solutions on the open source market as well as the commercial market. These editorials are to shed some light on the two most stable open source databases: MySQL and PostgreSQL. Both of these databases perform well and either can be used in a mission critical application. These editorials are just the opinions of the IM's that wrote them and are not endorsed by NSF or the LTER. I hope in forthcoming issues that someone will discuss the commercial products as well.***

What does PostgreSQL, the world's most advanced open source database, offer beyond the standard of an object-relational database management system?

- Sabine Grabner (MCR)

I am about to develop new information management facilities for MCR, one of the two newest stations within the LTER network, established in September 2004. My vision is to design and implement a relational database schema and related middleware and web-interfaces for inserting, updating and querying of datasets. On insert or update of data and metadata as content in the database, an application would generate or update a metadata document in EML with as little user interaction as possible or even automatically. The newest version of the EML document is automatically stored in the database and METACAT.

Since the most elegant and comprehensive collection of data is worth nothing once its meaning fades from the brains of their creators, it needs an even more elegant way to describe data and keep track of data manipulation, like EML. Especially in the context of our long term research goal, the value of our data is bound to its description. For my understanding it is also important to store the data itself in relational tables so that queries for future data demand or cross-site queries can be managed more easily than from static text files. The fact that an existing data format needs to be converted - a time-consuming task of writing parsing scripts - makes the use of those data very unattractive and makes the data management redundant in the first place. By using a RDB, data can be not only queried as a whole dataset, which context is somewhat arbitrary, but also by spatial, temporal or physical means.

For MCR's DBMS, I decided to use the open source object-relational database PostgreSQL. Besides its technical strength in the areas of administration, encoding, master-slave replication for distributed server architecture, backup, security, licensing, ACIDity and SQL standard compliance, and the fact that all this comes for free, PostgreSQL is also extensible, and this is one of its most attractive features which complements my goals.

Some extension packages come with the core PostgreSQL distribution, and many other extensions can be found at pgFoundry and GBorg or you can even design and implement new functions, operators, and custom data types yourself and add to the PostgreSQL server. The web sites also feature a whole list of software like graphical clients, packages to convert data and programs from other systems (such as MySQL, Oracle) into PostgreSQL, procedural languages, drivers, interfaces, database design tools, monitoring tools, administrator tools and developer tools.

The bitter part of open source software usually is the poor documentation, although this is not true for PostgreSQL. The online documentation is a straight forward reference, while the authors of the PostgreSQL guide book incorporated strongly their personal flavor and opinions, which is great for a newbie or if you've got some time to sit down and read as opposed to just pick up for one immediate simple answer.

So, what does this sweeping adjective "extensible" mean for storing, describing and accessing of very heterogeneous scientific data sampled by physical measurements, human observations or imaging, in varying temporal and spatial scale, as a function of time and/or space?

In the following I will cover three extensions which sound very promising for LTER information management purposes. xml2 and tsearch2 come in the core distribution and PostGIS is downloadable from http://postgis.refractions.net/index.php.

xml2 lets you execute XPath queries against XML documents stored in a PostgreSQL database and convert XML documents using XSLT style sheets. This feature allows storing metadata which are described by EML as a whole document and could eventually lead into a redesign of the database schema of METACAT for performance improvement. tsearch2 is a full-text indexing and searching package that lets you turn your PostgreSQL server into a search engine.

PostGIS adds support for geographic objects in PostgreSQL, allowing it to be used as a spatial database for geographic information systems (GIS), following the OpenGIS Simple Features Specification for SQL. Open source software compatible with a PostGIS enhanced PostgreSQL database as a backend, include GRASS GIS and MapServer (a development environment for building spatially-enabled internet applications).

Besides publicly available extensions such as these, you have the option of writing your own extensions (server-side programming) in Procedural Languages or High Level Languages such as C and C++. In the latter case simply compile your code into a dynamic object file (.dll or .so) and add it to the PostgreSQL server. The procedural language PL/pgSQL basically adds control structures such as conditionals, loops, and exception handling to the SQL language. As addition to server-side programming, PostgreSQL features a whole list of API's for client-side programming supporting a number of programming languages such as C, C++, Java, Perl, PHP, TCL, Python and many more.

In the following I will list some other notable features and important outlooks. Unique to PostgreSQL is GiST indexing, in addition to the supported indexes B-tree, Hash and R-tree. GiST stands for "Generalized Search Tree" which is a height-balanced tree structure and provides a possibility to create custom data types with indexed access methods. Unlike most other DBMSs PostgreSQL uses the multi-version concurrency control (MVCC) model to coordinate multi-user updates as opposed to locking a table against updates and queries during a single-user update. One of the improvements people from the PostgreSQL development team are working on right now is advancing the support of database server clustering.

This article on PostgreSQL, the world's most advanced open source database, is neither meant to point out any standard features a relational database management system RDBMS has to fulfill to comply the definition nor whether it is good or bad to use any non-standard features by means of portability of a schema. It is rather meant to point out some features which are unique to the open source client/server object-relational DBMS PostgreSQL with particular focus on those most useful to the needs of LTER information management. If I intrigued you on PostgreSQL, I would suggest both, the online reference and the PostgreSQL guide book for some further reading. If you are interested in a tabular comparison of DBMSs, wikipedia hosts an apparently up to date resource.

Experiences with MySQL 5.0 as a stand alone Relational Database in Ecological Studies

- Brian Riordan (BNZ)

Are you looking for a fast, reliable, easy-to-install and manage relational database? Do you need it to be free as well? Do you need tons of support options and training? Then MySQL 5.0 is the database for you. MySQL has all of the capabilities and then some to be the robust database you need for ecological applications and data.

The first MySQL database was developed May 23, 1995. In the past 10 years, MySQL has grown into a robust multithreaded, multi-user, RDBMS with an estimated six million installations. Many of the early criticism of MySQL was based on features that the MySQL team had not yet implemented into the database program. However, many of these features (Triggers, Stored Procedures, Row Locking) have been added with version 5.0.

MySQL offers a nice balance of reliability, security, and complexity while still remaining simple. Since MySQL is a company driven by a profit line, it offers many of the same features and support as the large commercial database systems. There is training and certification available. There are many levels of support an organization may purchase that go beyond the user forums and a large number of books. (A search on Amazon.com yields 210 resources for MySQL). There are typically large updates every year or two which strive to bring the database up to full compliance with ANSI/ISO SQL. There are also annual MySQL user conferences that provide hands on training with the developers as well as glimpses of what some users are doing with this database.

As MySQL was being developed they focused on what the creators call the 5-minunte installation. This means that they would like the user to be able to download their database and have a working version of it on their computer within 5 minutes. With MySQL I was able to be up and running within one day. In fact it is almost too easy to install as seen with issues concerning the default installation and the default users it creates. These problems have been correct in the most recent version.

For the BNZ - LTER the MySQL database has worked perfectly with our recent switch to Linux and Apache. We needed a relational database that could help us with more robust data mining and EML development. With so much support for the LAMP project MySQL was a logical solution. However, because the website was already coded in Coldfusion MX, we did not use PHP. Perhaps in the future we will explore a PHP migration.

MySQL develops many tools that help the DBA and database users interact with the databases. The three main tools that we use are:

- **MySQL Migration Toolkit:** A wizard driven database migration kit. This tool quickly moves an Access database into a MySQL database, making transitions between databases painless.
- **MySQL Administrator:** Offers an easy GUI for updating users, creating new tables, or managing security settings.
- **MySQL Query Browser:** An intuitive GUI for generating and optimizing SQL queries which is integral to a relational database and dynamic web coding.

In addition to the suite of tools that MySQL develops, there are a large number of programs (commercial and open source) available. We employ the use of several of these. Navicat and Toad are two of the commercial products we use to help manage tables and insert data.

The features that are missing in MySQL currently are easy to overlook in an ecological environment. This is especially true when many of the LTER sites look at their bottom line and current level of on site database support. In the case of BNZ, where we have one data manager that fills all of the IT needs, we need an easy, fast, well documented, free database. Comparisons between databases have shown that MySQL is the fastest database currently available. There are currently many large scale businesses (Yahoo, NASA, Ticketmaster, ect) that are moving their mission critical databases from the Oracles and Microsoft's over to MySQL. These large scale operations are employing MySQL and saving in some cases millions of dollars. Having said that, MySQL is not designed for the levels of referential integrity that a Postgres will provide. It often relies on application code to perform this critical check. This would be a huge problem to me if I had many different users updating my database. However, in my system I am the only Database user that can update or do "unique" queries. Anything added to the database is QC outside of the database as well as inside the database with row locking and cascading updates and deletes. If you have many users and are very concerned with referential integrity I might recommend one of the more robust databases. However, be aware that more database knowledge and planning will be needed.

I do not want to jump into specifics of the database (there are over a 100 books that can do that), but I do want to express that while MySQL is not Oracle, it is not trying to be either. It is a very fast database that has the ability to perform on a relation database level. There is a large amount of development and resources being poured into this database that will ensure it's future. In the 2 years that I have been working with the MySQL database I have witnessed vast improvements to security, speed, abilities, and support.

News Bits

Unit Dictionary Working Group Update

- Jonathan Walsh (BES)

I joined the Unit Dictionary Working Group in August, to help create online forms to enter units into the dictionary. I provide this status report based on my perusal of the emails to the group.

- One of the new units - milligramPerGramPerDay - forced the creation of both a new unitType and parentSI unit.
- UnitType expresses "what's being measured"... In the case of milligramPerUnitPerDay, that would be "MassPerMassPerTime". ParentSI expresses "what are the common SI units this can convert to" and in this case it's "KilogramsPerKilogramsPerSecond".
- The progress and all changes of the project are available through cvs.lternet.edu.
- Mark Servilla has a database for sharing and comparing custom unit definitions created by LTER sites on his server at http://fire.lternet.edu/customUnit/. If you simply search "all" you can get a good idea of what units have been defined.
- Wade Sheldon (GCE) has on his website, a metadata record for a data set that includes several custom unit definitions. It's visible online at: http://gce-lter.marsci.uga.edu/lter/asp/db/dataset_details.asp?accession=PHY-GCEM-0406a2 Click on the "Complete EML" link in the metadata section of the summary page and you can see how it looks expressed in EML.

The Unit Dictionary Working Group members are: Ackerman, Todd - NWT; Baker, Karen - PAL; Baudry, Jacques - INT; Costa, Duane - LNO; Elias, Pavol - INT; Garritt, Hap - PIE; Haber, Shaun - PAL; Millerand, Florence - PAL; O'Brien, Margaret - SBC; Powell, Linda - FCE; Ribes, David - PAL; Romanello, Samantha - LNO; San Gil, Inigo - LNO; Servilla, Mark - LNO; Sheldon, Wade - GCE; Walsh, Jonathan - BES; Yarmey, Lynn - CCE

Governance Working Group Proposes Updates to LTER By Laws

- Karen Baker (PAL/CCE)

Governance is the process through which a group makes decisions that direct their collective activity. The results of work by the LTER Governance Working Group (GWG) that was formed as part of the Planning Grant activities were presented at the 2005 March LTER Executive Committee meeting. The GWG presentation was titled "An Evolving LTER Network - Requires Changes to Governance". The group's charge was to consider modifications to the existing LTER Network structure to catalyze, embrace, and accommodate

-the conduct of synthetic inter site research on the grand scientific challenges,
-the constructive interaction with other research networks, and
-the growth of the Network that has and will continue.

GWG members present at the March06 meeting in Denver included Ann Zimmerman (Chair; Univ. Michigan, Social Scientist), Karen Baker (California Current Ecosystem and Palmer Station, Information Manager), Dan Childers (Florida Coastal Everglades, Lead PI), Chelsea Crenshaw (Sevilleta LTER, Graduate Student), Peter Groffman (Baltimore Ecosystem Study & Hubbard Brook, PI), Katherine Lawrence (Univ. Michigan, Organizational Behaviorist), John Magnuson (North Temperate Lakes, Interim LTER Network Chair), Larry Weider (Univ. Oklahoma, Organization of Biological Field Stations), and Bob Waide (LTER Network Office, Executive Director).

Specifically, the Executive Committee discussed changes that the GWG proposes to the LTER Network bylaws and the rationale behind those recommendations. One change is that the Executive Committee is being renamed to be Network Board. Another change in the by-laws is that there is a formalization of the participation of an information manager on the Executive Committee (that is, the Network Board).

The LTER By Laws section about the Information Management Representative currently reads as follows:

V 2.5 Information Management Representative *The LTER Information Management (IM) Committee shall select one member from among them to serve as the Information Management Representative to the Network Board. The IM Committee shall determine the method by which the Representative is selected. The Information Management Representative shall serve a single three (3) year term. In addition to the responsibilities shared by all Network Board members, the Information Management Representative shall serve as the liaison between the Network Board and the Information Management Committee providing insight on informatics, technology implementation, and human-technology infrastructure issues as well as on design and implementation of federated information system activities. In the event that an Information Management Representative is removed or not able to fulfill his or her term, the IM Committee will choose a replacement to complete the term.*

The proposed changes to the By Laws will be considered at the 2006 Spring Coordinating Committee Meeting.

Good Reads

Information Management at NTL: History, Evolution, and Insight

- Mark Servilla (LNO)

Long Term Dynamics of Lakes in the Landscape (J. J. Magnuson, T. K. Kratz, and B. J. Benson, editors) Oxford Press, New York. 2006.

The recent 2006 publication of ***Long-Term Dynamics of Lakes in the Landscape Long-Term Ecological Research on North Temperate Lakes*** by Oxford University Press includes a chapter on "Breaking the Data Barrier: Research Facilitation through Information Management." The chapter is organized around three central questions: (1) "How have information technology and management changed the way we do science?", (2) "What have been the pivotal decisions and principles that have shaped information management?", and (3) "What challenges does the future appear to hold for information management at North Temperate Lakes LTER?"

Analysis of the first question sets the motivation for solid information management beginning with an anecdotal experience where a study using historic research data (circa. 1930) on lake pH had to deal with the lack of documentation for analytical methods. Recreating the methodology required access to an oral history provided by members of the early research team - not the preferred method of documentation! The need for standards for data and metadata, coupled with the rapid advances in computer technology since 1980 (the start of NTL), led to the development of an information management infrastructure that promoted long-term legacy of information, with support for interdisciplinary research among NTL and other LTER researchers. The analysis also presents a broader vision for a network-wide information system that provides centralized discovery and access to distributed data by utilizing the latest tools that take advantage of semantic information embedded in metadata

The second question explores key decisions made with respect to NTL's information management philosophy. At the crux of this issue, NTL principal investigators decided from the beginning that core data will be collected and managed centrally - that is, the use of central resources and site-wide standards for data collection, documentation, and management were implemented at the start of the NTL LTER project. This strategy reinforced the development of information technology at NTL by defining two primary goals for the design of their information system: (1) "to create a powerful and accessible environment for the retrieval of information that facilitates linkages among diverse data sets" and (2) "to maintain database integrity." To achieve these goals, NTL utilized off-the-shelf software whenever possible to ensure industry standards and quality, while mitigating dependencies on tailor-made or in-house solutions, which are often rendered useless when the development team is no longer available for support. To streamline data management, NTL also is developing an automated QA/QC process that will significantly reduce the manual review of data from the sensor network. NTL recognized early on the need for an information management infrastructure that integrates computational science with field and laboratory based ecology. This is especially true, as information management at NTL employed both database and Internet technology for data archiving and access, and the on-going collection of geospatial data required extensive storage solutions and expertise in both GIS applications and remote sensing imagery analysis. As such, NTL required a dual-role information manager - part ecologist and part information technologist.

The third question addresses the future challenges of information management at NTL. Specifically, NTL faces both an increase in data volume from continuous monitoring of buoyed sensors and the vast array of Earth Observation imagery sensors, as well as new types of data that are generated by the social scientists who are now recognized as fulfilling an important role in long-term ecological research. These diversified data holdings, together with other distributed data sets, will act as the raw inputs for synthetically derived products that are only now being considered. For this reason, NTL sees the importance of educating new and emerging information managers through cross-disciplinary programs that can provide the necessary foundation to link information technology with ecological sciences.

From the challenges at the inception of the LTER program to a perspective on the future, this chapter provides a critical and historical review of the developmental process of the NTL information management infrastructure and can be viewed as a testament to the benefits of a planned information management system. As the Lead Scientist for the LTER Network Information System, I highly recommend this chapter as a "Good Read".

Designing Interfaces

- Shaun Haber (PAL/CCE)

J. Tidwell. Designing Interfaces. O'Reilly Media, 2005

We have all used plain, sterile, and sometimes ugly looking web applications before. These web apps may work well, but where's the fun in using them? Sometimes it's worth taking that extra step to give your application a "sexy" look. By doing so, the user receives an overall greater experience. A look at sites like Flickr and Gmail shows that not only are these great services, they also present visually pleasing and intuitive user interfaces.

In Designing Interfaces, author Jennifer Tidwell explores the many aspects of user interface design. This book is analogous to Design Patterns (see my other Good Read) in that it documents various interface design patterns, culminating in a wealthy repository of visual ideas for the interface designer. Additionally, Yahoo! has recently created its own online library of visual design patterns (http://developer.yahoo.com/ypatterns/index.php). These resources are extremely helpful for improving the front-end of any web application by promoting ideas for a more intuitive and aesthetic user interface.

Design Patterns: Elements of Reusable Object-Oriented Software

- Shaun Haber (PAL/CCE)

E. Gamma, R. Helm, R. Johnson, J. Vlissides. Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley, 1995.

Let's say we want to build a new web application for querying and analyzing ecological data. A service like this sounds pretty great, so where should we start? Assuming we already have a working infrastructure (web server, database, etc.), it's time to get dirty in the code. However, we should first answer a couple more questions: How do we best structure our code so to maximize code reuse and minimize duplicated efforts? What best practices and frameworks exist for promoting optimal code design?

Design Patterns is the first book to document object-oriented design patterns, providing programmers with a rich repository of reusable frameworks and preventing them from having to rediscover these solutions themselves. The four authors worked together to discover, analyze, and provide working examples of 23 previously undocumented design patterns. Essentially, they captured these patterns as abstract ideas, making it easier for developers to design, document, and share their code. (For the programmers, these patterns include the Flyweights, Factory Method, and Singleton class).

Although this book was written before the explosion of the web (it is geared for general software development), everything in it is applicable for building a web service. Be forewarned though: This book is fairly technical, and requires a background in computer science and/or object-oriented programming. That being said, with this book in our library, we can comfortably build any new software with efficiently designed code.

The Importance of Intertwingling

- Lynn Yarmey (CCE/PAL)

Morville, Peter. Ambient Findability . O'Reilly. Cambridge, 2005.

"Intertwingling" Aristotle and Edward O. Wilson, primitive hunter-gatherer societies and the Google experience, Peter Morville's Ambient Findability puts today's information age in the context of the vast human experience. Part philosophical treatise ('What is information?'), part history lesson (tracing the notion of a 'landmark' from the songlines of the aboriginal Australians through lighthouses and maps to a GIS marker) and part technology magazine (Subchapter Wayfinding 2.0 discusses specific SMS, RFID, and GPS technologies), he ties direct links between diverse facets of life and the design, usability and ramifications of technology and the internet. The many meanings of 'findability' become apparent as Morville blurs the line between the digital and the physical.

Many parts of this fast-paced book delve into the business end of information and it's architecture. Additionally there are chapters on such topics as language and representation, information retrieval, ontologies and 'The Social Life of Metadata' which certainly have import to the LTER community. Amongst other topics including economics and many mathematical considerations, the book also takes a close look at the human and social side of 'information,' expanding on the idea that technology alone cannot solve the challenges faced in today's information age. After a discussion on the nuances and development of language, including all of the ambiguity and personal perspective inherent to understanding, Morville notes, 'For as long as humans use language to communicate, information retrieval will remain messy, imperfect business.'

The text maintains it's lighthearted tone and easy-to-read style while giving an amazingly broad overview of our current information landscape, though unfortunately leaving little space to address specifics. Ambient Findability provides the larger picture and long-term path of the internet, from its grounding in basic human instincts to upcoming applications and uses, couched squarely in the middle of a complex and fascinating history of The User.

From Databases to Dataspaces: Opening up Data Processes

- Karen Baker (PAL/CCE)

Franklin, M., A. Halevy, and D. Maier, 2005. From Databases to Dataspaces: A New Abstraction for Information Management. SIGMOD Record 34(4): 27-33.

In the complex work of bridging data collection and data federation, our conceptual understandings grow regarding LTER research site data. These data efforts, informed and guided by local scientific needs and conventions, entail alignment of technological approaches and development of community nomenclature, standards, and dictionaries. Amidst such activities, this paper 'From Databases to Dataspaces' offers 'a new abstraction' that seems to escape the confines of a traditional data box model.

The authors are part of a group that has met over the last years to consider in depth the development, functionality, and use of data base management systems. This recent work opens up the data landscape conceptually - from databases to dataspaces. The 'dataspaces' approach is presented as both a new agenda and an architecture that allows for multiple ways of solving issues and framing questions of information management. In addition, development is recognized to take place over multiple timeframes: "One of the key properties of dataspaces is that semantic integration evolves over time and only where needed. The most scarce resource available for semantic integration is human attention." The dataspace concept umbrellas explicitly a continuum of organizational and semantic arrangements that handle diverse data types, states, and approaches. The paper recognizes the multiple facets of work involved: "Dataspaces are not a data integration approach; rather, they are more of a data co-existence approach". In the paper, there is discussion of dataspace requirements, components, and research challenges.

Perhaps this paper struck a chord because it seems to resonate with one of my first impressions of LTER, that is, a community recognizing the value of a good number of distinctly unique LTER sites working in loose proximity. For LTER, years of joint projects addressing local, regional, and cross-site science have contributed to a thick infrastructure. This infrastructure includes a shared sensitivity toward respect for local diversity and trust that new knowledge as well as research challenges arise from grappling with heterogeneity. As we seek tighter collaborative configurations in both ecological science and informatics endeavors, the dataspaces concept highlights the notion of a development process over time of data and information management, of theory and practice.

Alaska's Changing Boreal Forest

- Brian Riordan (BNZ)

Chapin, F.S., III, M.W. Oswood, K. Van Cleve, L.A. Viereck, and D.L Verbyla (editors). 2006. Alaska's Changing Boreal Forest. Oxford University Press, New York

The Bonanza Creek LTER released its synthesis volume this spring: Alaska's Changing Boreal Forest. This volume synthesizes current understanding of the ecology of Alaska's boreal forests and describes their unique features in the context of circumpolar and global patterns. It tells how fire and climate contributed to the biome's current dynamics. As climate warms and permafrost (permanently frozen ground) thaws, the boreal forest may be on the cusp of a major change in state.
This synthesis volume covers a wide range of topics and research prevalent to a warming climate and it's impacts on the Boreal Forest.

In the first section, the reader can absorb the geographic and historical context for understanding the boreal forest. The book then delves into the dynamics of plant and animal communities inhabiting this forest, and the biogeochemical processes that link these organisms. The last section explores landscape phenomena that operate at larger temporal and spatial scales and integrates the processes described in earlier sections. Much of the research on which this book is based results from the Bonanza Creek Long-Term Ecological Research Program. This book provides a synthesis of the substantial literature on Alaska's boreal forest that is accessible to professional ecologists, students, and the interested public.

In summary, recent warming has increased drought stress in trees, frequency of wildfire and insect outbreaks, sporadic loss of permafrost, loss of nitrogen from nitrogen-limited watersheds, and changes in feedbacks to regional and global climate systems. Many of these changes have substantial societal importance both regionally and globally, suggesting that continuing efforts to develop a predictive understanding of ecological resilience and change are critical.

As with all of the LTER synthesis volumes this is a "good read".

Frequently Asked Questions

USGS streamflow data automatically harvested for inclusion in HydroDB/a>

- Wade Sheldon (GCE)

Question: How can our site sign up to have USGS streamflow data automatically harvested for inclusion in HydroDB?

Answer: First you need to identify candidate real-time streamflow stations near to your site using the USGS station search forms at http://waterdata.usgs.gov/nwis/rt. Next, send your list of station ids to Suzanne Remillard (AND) , and she will create corresponding HydroDB station ids and forward the list to Wade Sheldon (GCE) for testing. An initial harvest will then be performed, and you will receive a table listing the date coverages, available parameters, and parameter value ranges for each of your stations. If data values at any of your candidate stations regularly exceed the default HydroDB quality control limits (e.g. discharge > 20,000 lps), you may need to use the ClimDB/HydroDB metadata forms to specify higher limits. After your sites are verified and metadata records are updated, automatic data harvests will commence weekly and you will receive harvest summaries via email. (For more information on this service, see http://gce-lter.marsci.uga.edu/lter/research/tools/usgs_harvester.htm).