LTER

01001100 01010100 01000101 01010010

LTER DataBits

Information Management Newsletter of
The Long Term Ecological Research Network

01001100 01010100 01000101 01010010

Feature Articles

News Bits

Good Reads

Calendar

DataBits:An electronic newsletter for Information Managers ----- Spring 2005 Issue (http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/05spring/)

Featured in this issue:

Following years of efforts in developing a Network Information System (NIS), LTER information managers report on two of its completed components: CLIMDB/HYDRODB and Network All-Site Bibliography. More recently, the LTER sites started the process of exporting their metadata databases into Ecological Metadata Language (EML), a "metadata specification developed by the ecology discipline and for the ecology discipline" (http://knb.ecoinformatics.org/software/eml/). This task has proven itself to be a complex enterprise, whose ramifications cover a wide variety of data management activities and issues. In this article, we can perceive part of this scope as several articles cover from an EML-related tool providing a harvesting service to the development of shared dictionary template. A new harvesting service developed by the LNO that provides LTER sites with means of synchronizing metadata documents with the LNO/KNB Metacat server and EcoGrid networks is presented. A summary of the survey on the resources required by each LTER research site to implement a strategy for generating EML is given. The committee working on identifying Web best practices give their first report. A group of information managers that emerged to discuss the need and peculiarities of developing a dictionary "as a mechanism for moving toward interoperability of site data and cross-site data" give their first insights. Another feature presents a comprehensive graphical user interface (GUI) search engine application for building and managing search indices, defining queries, and managing result sets from searches as developed by GCE. Finally, as a reminder that we all belong to a wider community, constantly facing all kinds of challenges, including the most ancient in human history, facing Nature, a group from the NTL site tells us a story on how they survived a typhoon while installing wireless sensors in Taiwan's Yuan Yang Lake.

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: Eda C. Meléndez-Colom(LUQ), Jonathan Walsh(BES)

Feature Articles

EML Harvesting II: Preparing Site Metadata and Harvest Lists

- Wade Sheldon (GCE)

INTRODUCTION

In **EML Harvesting I** (DataBits, Fall 2004), Duane Costa described the features and operation of the new Metacat Harvester software and associated harvesting service developed at the LTER Network Office (LNO). This new harvesting service provides LTER sites with a simple and practical means of synchronizing metadata documents with the LNO/KNB

Metacat server, and by extension the broader Metacat and EcoGrid networks. This Harvester specification has also recently been adapted by Chris Lindsley and Tim Rhyne at the Oak Ridge National Laboratory (ORNL) to support automatic harvesting of EML metadata and transformation to FGDC Biological Data Profile (BDP) format for inclusion in the NBII Metadata Clearinghouse.

Thanks to these dual efforts, LTER sites can now participate in multiple metadata search networks by:

1. Providing valid EML documents (compatible with Metacat) on a publicly-accessible WWW server
2. Creating an XML harvest list document containing the WWW URL for each EML document
3. Registering the URL of the harvest list with LNO and scheduling harvests (and optionally registering for NBII participation)
4. Managing document "revision" numbers and corresponding EML "packageIDs" to control the harvest and Metacat synchronization process as site EML documents are added or revised

These steps are described in the remainder of this article.

1) Providing EML for Metacat

The KNB Metacat data repository system is designed to archive XML-based metadata documents regardless of their schema, and Harvester is similarly schema-agnostic. The only nominal requirements are that documents conform to XML structure rules (i.e. are well formed), and are valid according to the referenced schema. In the specific case of EML, this means document contents must conform to the EML 2.0.0 or 2.0.1 schema rules as documented at http://knb.ecoinformatics.org/software/eml/.

However, the following guidelines taken from the EML Best Practices for LTER Sites document must also be followed in order to support automatic document harvesting and synchronization with Metacat:

a) EML document ids and revision numbers:

"packageId" attributes for EML contributed to the KNB Metacat should be formed as follows:

<eml:eml packageId="knb-lter-[site].[dataset number].[revision]" system="knb" ...> (e.g. packageId="knb-lter-gce.187.4")

b) Access Control:

Metacat access control format conforms to the LDAP distinguishedName for an individual, as in “uid=FLS, o=LTER,dc=ecoinformatics,dc=org” (where "FLS" stands for "Fictitious LTER Site"). Access elements for documents contributed to the KNB Metacat should be formed as follows:

```
<access authSystem="knb" order="allowFirst" scope="document">
   <allow>
      <principal>uid=FLS,o=lter,dc=ecoinformatics,dc=org</principal>
      <permission>all</permission>
   </allow>
   <allow>
      <principal>public</principal>
      <permission>read</permission>
   </allow>
</access>
```

Specific access control rules can also be included for any individuals registered in the KNB LDAP server, such as the site IM or contributing PI; however, LNO has established an alias account for each site (based on the three letter site acronym, e. g. uid=GCE) to ensure consistent ownership of LTER metadata stored in the KNB Metacat independent of personnel changes over time. Site IMs can contact Duane Costa <dcosta@lternet.edu> to obtain or reset the password on their site alias account.

2) Creating a Harvest List Document

The Metacat Harvester operates by periodically downloading and parsing an XML-based "harvest list" document containing URLs for all EML documents available at the harvest site. This site-managed harvest list is therefore the key to participating in the Metacat EML harvesting system, much like the legacy LTER DTOC cataloging system.

The structure of the harvest list is fairly simple, as illustrated in the following example and fully described on the new LTER IM Mentoring web page. A "<document>" element is required for each EML document to be harvested, listing a unique numeric identifier, numeric revision number, and WWW URL. As in the example, URLs for both static EML documents and web applications or scripts with query string parameters can be included as appropriate. Note that any XML reserved characters in URLs, such as "&", "<" and apostrophes, must be "escaped" using the XML character references "&", "<" and "'", respectively.

```
<?xml version="1.0" encoding="UTF-8"?>
<hrv:harvestList xmlns:hrv="eml://ecoinformatics.org/harvestList">
  <!-- first EML document -->
  <document>
    <docid>
      <scope>knb-lter-gce</scope>
      <identifier>1</identifier>
      <revision>7</revision>
    </docid>
    <documentType>eml://ecoinformatics.org/eml-2.0.1</documentType>
    <!-- static EML document URL -->
    <documentURL>http://gce-lter.marsci.uga.edu/lter/datasets/eml/knb-lter-gce_1_7.xml</documentURL>
  </document>
  <!-- second EML document -->
  <document>
    <docid>
      <scope>knb-lter-gce</scope>
      <identifier>2</identifier>
      <revision>6</revision>
    </docid>
    <documentType>eml://ecoinformatics.org/eml-2.0.1</documentType>
    <!-- dynamically-generated EML document URL -->
    <documentURL>
      http://gce-lter.marsci.uga.edu/lter/asp/db/send_eml.asp?detail=full&amp;
missing=NaN&amp;metacat=yes&amp;dataset=2
    </documentURL>
  </document>
  <!-- additional EML document elements... -->
</hrv:harvestList>
```

Although the harvest list structure looks somewhat verbose, most of the document is composed of static markup. The only variable portions are the two numeric fields and the URL itself (highlighted in red in the example code). The harvest list

can therefore be generated very easily by cutting and pasting in a text or XML editor, or using simple string handling procedures in any scripting language.

3) Scheduling Harvests

After sites have posted valid EML documents to a WWW server and have constructed a corresponding harvest list, the URL for the harvest list and harvesting frequency must be registered at LNO as described in **EML Harvesting I** (DataBits Fall 2004). The ideal harvesting schedule for a site will depend on the frequency with which the site typically adds or updates data sets and metadata, and makes corresponding changes to revision numbers in the harvest list. Monthly or weekly harvests are probably reasonable for most sites, although more frequent harvests could be requested considering that only new or updated documents will be retrieved so system resources will not be needlessly taxed on either end.

In order to register for NBII harvesting, site IMs should individually contact Inigo San Gil <isangil@lternet.edu> or Tim Rhyne <rhynebt@ornl.gov> for assistance; however, this policy may change in the future because Inigo will be investigating the possibility of NBII harvesting metadata directly from the LNO/KNB Metacat server. After sites are scheduled for harvesting, NBII personnel will follow up with the site IM to request general information for creation of a "Clearinghouse Node" description for their LTER site. For sites that wish to further advertise their data holdings, NBII can also publish their metadata in the FGDC clearinghouse (also called the National Spatial Data Infrastructure [NSDI]) and the new Geospatial One-Stop (GOS) on request.

The same harvest list URL can be registered for Metacat, NBII, NSDI and GOS participation, or separate URLs can be registered to specifically tailor the metadata documents synchronized with each system. At GCE, for example, we provide complete harvest lists for Metacat, NSDI and GOS synchronization, but generate a reduced harvest list containing only URLs for metadata from biologically-oriented studies (based on research theme) for NBII. These harvest lists are dynamically generated from a single web application (http://gce-lter.marsci.uga.edu/lter/asp/db/eml_harvest_doc.asp), using a query string parameter to distinguish among synchronization targets (i.e. hostname=metacat for Metacat, hostname=nbii for NBII and hostname=geospatial for NSDI and GOS); however, multiple static harvest list documents could also be produced to accomplish the same task.

4) Managing EML Harvests

As indicated in the *Providing EML for Metacat* section, Metacat and by extension Harvester rely on numerical data set ids and revision numbers for document management and synchronization. When Harvester encounters a new *<identifier>* or changed *<revision>* in a harvest list, the corresponding EML document will be downloaded and inserted into Metacat; consequently, sites can only control metadata harvests by managing these identifiers and revisions. Although this sounds straight forward, theoretical and practical issues concerning data and metadata versioning have been hotly debated in LTER for many years and versioning practices vary extensively across site information systems. Even for sites that do number and version data sets, work-arounds may be required for maximum interoperability with Harvester and Metacat.

At GCE we use sequential numeric ids as alternative identifiers for all data sets in our metadata database and we maintain explicit major and minor version numbers to track changes in data and metadata content since original release. Despite this apparently idyllic situation for Metacat compatibility, we had to devise a complex work-around for generating EML revisions to accommodate changes in EML implementation independent of metadata contents. For instance, we made several changes to our EML implementation in February-March 2004 in response to feedback from NCEAS developers and to improve display of our documents using the default Metacat XSL style sheets. Further changes were prompted by the EML Best Practices working group meeting in May 2004. Each of these changes required a revision change in order to trigger re-harvesting of the updated EML documents despite the fact that the underlying data and metadata contents themselves had not changed.

The best strategy for supporting Metacat versioning (as well as the KNB authentication system) in EML documents will likely vary according to the technology used to generate the EML documents themselves. Sites that plan to manage static documents may have to manually update and synchronize revision numbers between EML documents and the harvest list or develop scripting or XSLT approaches to propagate version changes. Sites that manage metadata in an RDBMS and generate EML documents and the harvest list programmatically may choose to add EML revision tracking to their systems, or just periodically increment revision numbers to force updates in Metacat. At GCE we have taken this process a step further and chosen to differentially generate EML optimized for Metacat. Document URLs in the dynamically-generated GCE harvest list contain an additional query string parameter "&metacat=yes", which instructs our web application to include Metacat-specific packageIds and revisions, appropriate access control elements for the KNB system, and alternative data table URLs designed to stream data for publicly-accessible data sets (i.e. after transparently logging access in our data use tracking system). Many versioning strategies and work-arounds are possible and LTER Information Managers are encouraged to discuss potential approaches with LNO, the EML Best Practices working group, and other IMs as they develop support for EML harvesting at their site.

Note that NBII completely replaces all metadata records with the current ones during each harvest; therefore, versioning issues are not critical for NBII, NSDI, and GOS harvests. Sites can also request a special off-schedule harvest if major changes are made to their EML implementation or documents. Consequently, managing harvests for NBII, NSDI and GOS participation is considerably simpler than for Metacat.

CONCLUDING REMARKS

The Metacat Harvester fills an important technology gap that has prevented many LTER sites from participating in the Metacat repository system. Although the KNB Morpho program is a powerful metadata entry and management tool that works directly with Metacat, the lack of built-in support for metadata content re-use and integration with existing site information systems has precluded its use at most LTER sites. Other systems capable of synchronizing metadata with Metacat (e.g. CAP-LTER Xanthoria) have also not been adopted by most sites for various technical reasons.

The technological neutrality of the Metacat Harvester is particularly beneficial from a site perspective, because it supports participation regardless of IT architecture or EML generation approach and will accommodate transitions in technology over time. For example, a site just beginning to generate EML metadata can maintain a static harvest list and update the list in a text or XML editor as each new document is created, whereas a site developing more automated approaches can generate a dynamic harvest list using any web application framework. Similarly, URLs for both static and dynamic EML documents can be included in a single harvest list, allowing sites to develop dynamic EML-generation capabilities in stages without affecting participation in metadata search networks.

Contributing LTER EML documents to the LNO/KNB Metacat will help towards accomplishing a major goal identified by the LTER NIS Advisory Committee, by providing integrated data searching across the LTER Network based on structured metadata. It will also allow LTER sites to leverage tools and technologies being developed by KNB and SEEK built on Metacat, such as the EcoGrid and Kepler work-flow analysis software. The ability to synchronize metadata with the NBII Clearinghouse with no additional effort is also a tremendous benefit to both the LTER and NBII networks, and will support discovery and use of LTER data by an even wider audience in the scientific community.

ACKNOWLEDGEMENTS

I would like to thank David Blankman (formerly at LNO), Duane Costa (LNO), Matt Jones (NCEAS), and Chris Lindsley (ORNL) for their collaboration and help developing support for EML harvesting at GCE, which led to this article. I would also like to thank Tim Rhyne (ORNL) for providing additional information on NBII and for his editorial advise.

Evaluating First Generation LTER Site Web Sites: Assessing our audience, meeting their needs, and making recommendations for the future.

- Nicole Kaplan (SGS), Corinna Gries (CAP), and Eda C. Meléndez-Colom (LUQ)

The first Long Term Ecological (LTER) sites were established in 1980, and in the 1990s existing sites seized web technology to publish information about their research, sites, data, personnel, publications and opportunities to the ecological community and general public. Over the past decade the National Science Foundation and the LTER Information Management (IM) community has addressed expectations for web accessible network-wide data and information, data release and access policies, and recommendations for tracking data usage. Since our first generation LTER web sites were launched, our users' expectations, information technologies, and design techniques have evolved. Our challenges today are serving gigabytes of metadata and data for hundreds of data sets from a federated system of LTER web servers, updating backend databases with dynamic content, and keeping up with drifting standards and media in which to publish data in various formats.

In a recent survey, LTER sites identified their web site audience, reported on how they tracked usage of information from web sites, and evaluated the effectiveness of various web site components. Twelve sites responded to the survey and the results suggest that LTER sites that have created successful web sites went through a process to identify their user groups and included input from those groups while developing web sites and tools. Sites measure the success of their web sites by assessing whether people can easily find usable information, data and metadata. As web users (and reviewers), we tend to measure the success of a web site by the number of clicks it takes to query and reach downloadable, well-described, and useful data and metadata. More detailed survey results and links to successful LTER web site tools may be found here: http://intranet.lternet.edu/committees/information_management/im/webdev_questions_summary.pdf.

Based on the experiences with first generation web sites, several challenges have been identified. Web tools and web sites have been developed independently at each LTER site. Therefore, each has a different look and feel reflecting the individuality of each site within the network. Categories of information are not well-defined and do not contain consistent information across sites. A user finds different content arranged under various categories from web site to web site. Data are updated at irregular intervals, and may be presented as screen shots or in a downloadable format. Data are not necessarily related to projects or publications. Publication of news bulletins, exciting new findings and products is handled differently at each site and are generally difficult to find. Finally, LTER web pages do not have design elements, links, or text to portray themselves as being part of a larger network or connect to other LTER sites easily.

A working group was formed within the LTER IM community to address these challenges. The group plans to create recommendations for developing new or second generation web sites that will allow sites to maintain their own web tools that fit in with their local organizational structure and meet the needs of their local site users. The working group is currently developing LTER web content guidelines, considering a controlled vocabulary for keywords, under which the datasets may be found (http://knb.ecoinformatics.org/index.jsp), and elements and links for homepages that may create the feel of a network identity across LTER web sites. A workshop will be organized at the 2005 Annual IM Meeting in Montreal to discuss recommended web design techniques to make navigation of individual sites and connections to network cohorts easier. Approaches to get more feedback from web site developers and users on the successfulness of our LTER web sites also will be discussed.

Here are some useful resources recommended by Marshall White, LNO Web Developer, for designing web sites that meet the needs of their users and incorporating good design components:

Web Design in a Nutshell, 2nd Edition
http://www.amazon.com/exec/obidos/tg/detail/-/0596001967/104-1363550-8711147?v=glance

Information Architecture for the World Wide Web
http://www.amazon.com/exec/obidos/tg/detail/-/1565922824/104-1363550-8711147?v=$

Customer-Centered Design: A New Approach to Web Usability
http://www.amazon.com/exec/obidos/tg/detail/-/0130479624/qid=1105646463/sr=8-1/ref=sr_8_xs_ap_i1_xgl14/104-1363550-8711147?v=glance&s=books&n=507846

Usability for the Web: Designing Web Sites that Work
http://www.amazon.com/exec/obidos/tg/detail/-/1558606580/ref=pd_sim_b_2/104-1363550-8711147?%5Fencoding=UTF8&v=glance

GUI Bloopers: Don'ts and Do's for Software Developers and Web Designers
http://www.amazon.com/exec/obidos/ASIN/1558605827/qid%3D1105646210/sr%3D11-1/ref%3Dsr%5F11%5F1/104-1363550-8711147

ClimDB/HydroDB (ClimHy) Database Update

- Suzanne Remillard and Don Henshaw (AND)

The Climate and Hydrologic Database Project is a component of the LTER Network Information System (NIS), the suite of LTER intersite database modules being created to promote synthetic ecological research. The ClimDB and associated HydroDB were developed in response to research scientist need for current and comparable climate and hydrologic data summaries for LTER sites and U.S. Forest Service Experimental Watersheds widely used in intersite comparisons, modeling studies, and land management-related studies (Henshaw et al. 1998). ClimDB is composed of a harvester system that continually captures and updates data from multiple sites into a central database, as well as a web interface that allows graphical display and download of data in common formats (Baker et al.2000).

Numerous modifications have been made to the ClimDB/HydroDB databases (affectionately referred to as ClimHy) over the past two years. Visual modifications have been made to both the public and participant web interfaces, and many behind-the-scene revisions have improved the functionality for the database users. Please note the new ClimHy URLs:

Participant Page - http://www.fsl.orst.edu/climhy/harvest/harvest.htm

Public Data Access Page - http://www.fsl.orst.edu/climhy/

The earlier web URL's redirect users to these new pages. Additionally, both of these web page URL's are posted on the LTER network intranet.

One of the major changes is the integration of the ClimDB (primarily NSF-funded) and HydroDB (primarily USFS Research-funded) web pages. This integration has created a better integrated, one-stop shopping portal for climate and hydrology data and metadata. The data has always been stored in the same database, but the databases had

unique access points from the web. One major enhancement to the Public Data Access Page is a handy interactive table for viewing the sites, stations and variables included in the database by organization: LTER, USFS, or USGS. Once displayed, the sites, stations and variables can be sorted and re-listed by any displayed field including by the last harvest date or by the most current data harvested. In addition, the "Data, Plots, and Downloads" allow viewing data as html, downloading as comma or tab-delimited files, or viewing graphical displays that can be saved as image files. Graphical displays are also improved and include plots of one site versus another or meteorological versus hydrological data, and long-term aggregated means can be displayed alongside monthly or annual data. Finally, we now dynamically write a PDF that includes complete site metadata for all stations in a single report.

The most recent enhancement is the initiation of our auto-harvest feature, which will initiate a harvest weekly (currently Sunday mornings). Several sites have already written scripts to automatically harvest, but sites interested in participating in the weekly auto-harvest feature, please contact the ClimHy Data Manager (climdb_admin@fsl.orst.edu). Sites can trigger a harvest as a scheduled event or through a program by using the following URL:

http://wwwdata.forestry.oregonstate.edu/climhy/harvest.pl?module=<#>&site=<ltersite>

where,
<#> = URL option number (1 or 2)
<ltersite> = Three letter LTER Site code (i.e., AND for Andrews)

Two URL's are maintained in the database for each site (option 1 is the previous ClimDB URL, option 2 is the previous HydroDB URL), but either URL can be used for any data set or scripting program (option 3 is reserved for USGS data harvest only). These URLs can be edited from the participant page under "Update Metadata" and then "Research Area Information". Sites employing real-time USGS Gauging Station measurements can have this data regularly harvested (LTER Network News, Fall 2003). Please contact the ClimHy Data Manager.

The User Guide has been updated and a FAQ is planned. Many of the pages feature a feedback form, which allows the user to submit comments, questions, or suggestions about the various modules or applications. A data access policy has been established with a data use agreement, disclaimer and general citation. Data users are encouraged to consult or collaborate with original investigators, and a link to site contact personnel is provided. ClimHy will adopt the new LTER Data Access and Use Policy, once approved.

What about web services for ClimHy? San Diego Supercomputer Center scientists and LTER Information Managers demonstrated web services architecture and successfully built a prototype involving a few LTER sites (LTER Network News, Fall 2002). While existing architecture has not been replaced for harvesting individual sites using web services, the entire ClimDB/HydroDB database can now be programmatically accessed through web services. Two services allowing access to both raw and aggregated data are described and can be tested from the following URL: http://wwwdata.forestry.oregonstate.edu/climhy/climhy_ws_api.htm

The ClimHY resource continues to grow and increase in use. There are now 6.2 million daily values in the database. All ClimHy web sessions, data views, downloads, plots and feedback are logged into database tables for tracking purposes, and the site is monitored by WebTrends software for general usage statistics. The public web page is averaging 28 visitor sessions per day for the first two months of 2005 compared with 20 sessions per day throughout 2004, with nearly 4000 unique visitors and over 1000 files graphically displayed or downloaded in the past 14 months. Thank you to all contributors who continue to make the ClimHy enterprise a successful venture.

References:

Baker, Karen S.; Benson, Barbara J.; Henshaw, Don L.; Blodgett, Darrell; Porter, John H., and Stafford, Susan G. Evolution of a multisite network information system: the LTER information management paradigm. BioScience. 2000; 50(11):963-978. http://www.fsl.orst.edu/lter/pubs/webdocs/reports/pub2742.pdf

Henshaw, Donald L.; Stubbs, Maryan; Benson, Barbara J.; Baker, Karen; Blodgett, Darrell, and Porter, John H. Climate database project: a strategy for improving information access across research sites. In: Michener, William K.; Porter, John H., and Stafford, Susan G., eds. Data and information management in the ecological sciences: a resource guide; 1997 Aug 8-1997 Aug 9; Albuquerque, NM. 1998: 123-127. http://intranet.lternet.edu/archives/documents/data-informationmanagement/DIMES/html/henshaw2.fv2.htm

Henshaw, Don; Sheldon, Wade; Vanderbilt, Kristin. Introducing the Climate and Hydrology Web Harvester System. The Network Newsletter Fall 2003; 16(2):8. http://intranet.lternet.edu/archives/documents/Newsletters/NetworkNews/fall03/fall03_pg08.html

Vanderbilt, Kristin. SDSC and LTER: Demonstrating a Web Services Architecture. The Network Newsletter Fall 2002; 15(2):9. http://intranet.lternet.edu/archives/documents/Newsletters/NetworkNews/fall02/fall02_pg09.html

---

The LTER Network All-Site Bibliography

- James W Brunt, LTER Network Office (LNO)

PROJECT DESCRIPTION

The U.S. Long-Term Ecological Reseach (LTER) Network All-site Bibliography serves to account for the scientific contributions of the LTER Program, facilitate cross-site synthesis and synthetic studies, and generate new interest in LTER sites and LTER research. The infrastructure supporting the bibliographic records and the functionality of it's interfaces continue to evolve as described in Brunt and Maddux (2002). There have been some major jumps forward – each time producing a functional new product. However, until now, none of these products have been maintainable and thus quickly became obsolete. The first successful attempt used unique delivery scripts for each site that quickly became obsolete. This solution also relied on indexing software that, you guessed it, quickly became obsolete. The task of building a distributed all-site bibliography was intractable because of the heterogeneity of the way sites were storing and managing their bibliography data and the frequency at which these methods changed. More recently, we standardized on a particular end-user software package, EndNote®, to take advantage of a proprietary web-publishing solution, Reference Web Poster®. This solution was successful because of the standardization but extremely limited in the way information could be retrieved and used. Neither of these previous solutions had the power of an open relational database management system behind them. This was the most important requirement in attempting to provide any value-added components to this very useful data set.

The LTER All-Site Bibliography has now hopefully completed its last platform migration. Now, finally, implemented in a highly normalized relational database model, it should serve the LTER Network for years to come. Moving the bibliography database out of proprietary software and into an open rdbms framework has allowed us to now focus on generating standards-based outputs, easing the burden of updates for sites, providing useful searching and reporting, and providing virtual bibliographies.

REQUIREMENTS

The LTER Network Information System (NIS) working group, in 2002, established the following requirements for the All-Site Bibliography database and interface improvements:

1. Must be able to accept EndNote export format as input
2. Must be able to distinguish duplicates from new entries and provide appropriate administrative responses
3. Must be able to house multiple media types and distinguish between them
4. Must be able to provide site lists and individual lists based on authentication
5. Must be able to produce endnote export format
6. Must be able to provide personalized lists
7. Must have compatible elements to satisfy Ecological Metadata Language (EML) literature
8. Must have z39.50 connection for compatibility with bibliography clients

These requirements have now all been met with the implementation described here.

DATA MODEL

The database model is based on the EndNote generic type with additional attributes for LTER specific content and additional tables to manage authors names and keywords at a higher level of normalization. The model was adapted from one in use at CAP LTER.

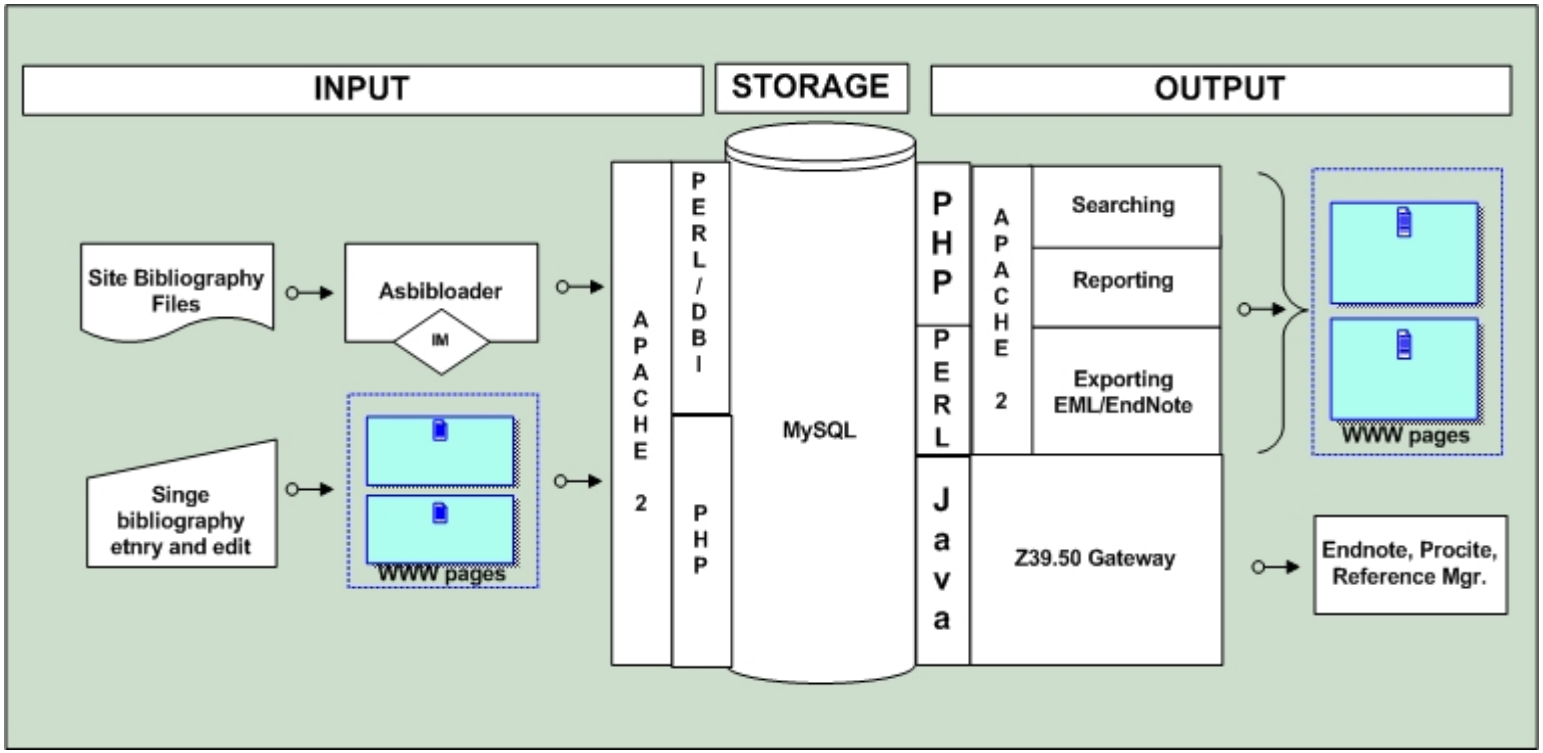

Figure 1-1. LTER All-Site Bibliography Architecture

In 2002 we mapped out an architecture for the future of the LTER All-Site Bibliography (Brunt and Maddux, 2002). Although all components of the architecture were not specified we have followed this architecture as a roadmap for development. Figure

1-1 shows the current architecture as it is implemented. The 26,000 bibliographic data records are now housed in a MySQL database management system instance running on MS Windows 2000 Server®.

SEARCHING

There are currently two interfaces for searching, web via php (http://search.lternet.edu/biblio/) and Z39.50 (See Box 1-1; Note: screen shots of the interfaces are not included here because they are at present too dynamic). The php scripting is implemented under an Apache 2 web server running on RedHat Linux ES. These scripts set a PHPSESSID variable on the local browser to handle the marking and exporting of subsets of bibliographic entries. A simple search that searches on title, author, and keyword fields and an advanced search that offers some limited boolean searching on specific fields are implemented. The EndNote generic type information is automatically converted in the search interface and displayed based on the reference type attribute in the web search and administrative interfaces. For example, a journal article will display the word "Journal" instead of the generic type attribute "Secondary Title".

Box 1-1. - To connect to the Z server from EndNote:

1. Install the attached LTERBIB.enz file in the appropriate directory, Eg., C:\Program Files\EndNote\Connections directory (or equivilent if installed elsewhere)
2. Go to the "Tools" pulldown menu in EndNote. Select the "Connect..." option.
3. Scroll alphabetically through the options there until you find "LTERBIB". Select this entry and click "Connect".
4. Explore
5. Provide Feedback to me. If you get stuck or have questions about the data you find there drop an email to tech_support@lternet.edu.
6. Known Bugs
    1. One of the bugs found so far is that that any Boolean 'AND' that involves 'Year' doesn't work correctly - it seems to be consistent in it's errors but I haven't tracked them down yet.
    2. A few of EndNotes fields don't make into the Search Client mapping appropriately - one for sure is "Short Title"

The Z39.50 compliant server is implemented in java as a swing application and uses an XML serializer and xalan stylesheet processor to map between the database fields and the Z39.50 map (Figure 1-2). Z39.50 is widely used by the library community and is based on the dublin core metadata standard. Which means you can now search the bibliography via Z39.50 compatible clients. We specifically mapped this server to the EndNote generic type to make it more useful as a connection client tool in EndNote. Once the references are downloaded their document type can be changed in EndNote for better display. Z servers are not terribly fast, so if you do a large search and retrieval it could take a while. In this prototype, we have not attempted to optimize for speed of delivery.

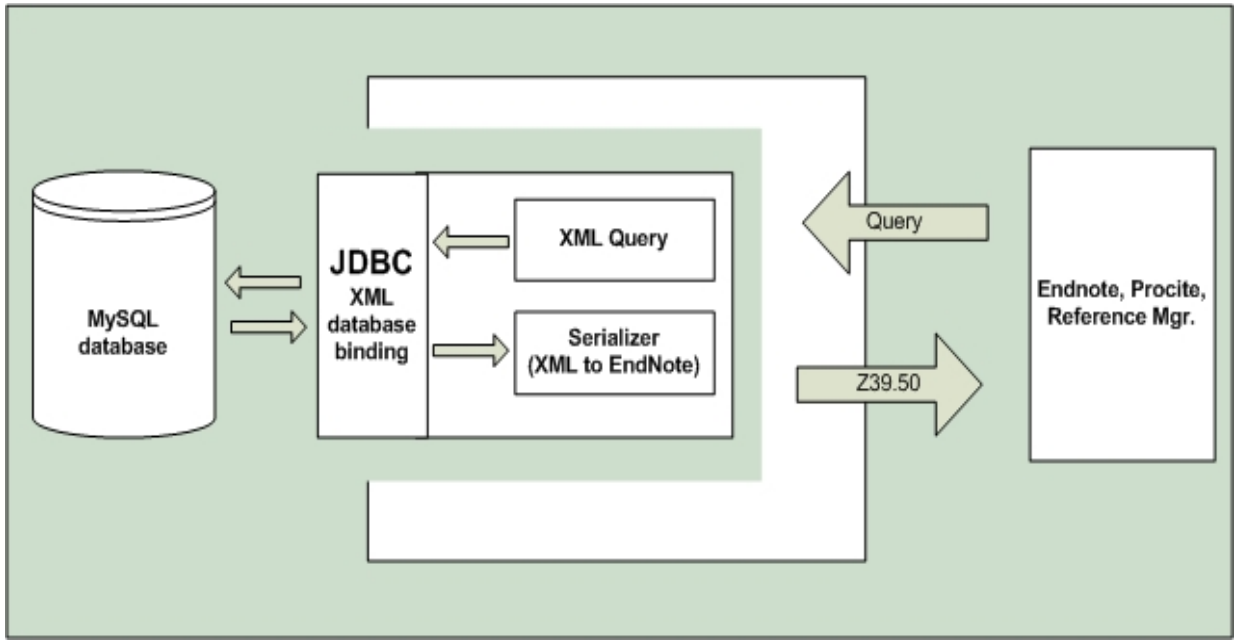

Figure 1-2 – LTER All-Site Bibliography Z39.50 Gateway

Most of the common fields contained in the bibliography database have been mapped to the Z server and back out to the EndNote fields again. Some improvements on searching, allowing searching by Site Code and Reference Type (Journal Article, Book Section, etc.) have been made in this release. Even though these fields don't normally show up in EndNote it is possible through the miracle of XSLT to query them anyway and return the EndNote formatted records You can now get all the Journal Articles in one query. You can also make them show up as EndNote Journal articles if you so desire by editing the connection file (under the edit pull down) and resetting the "default reference type" to Journal Article. Yes, kind of a round about approach but unless it's a big search it's not that difficult to change them and the citation formatting seems to know what to do with them even if they are generic.

ADMINISTRATION

The administrative interface is also php based but calls perl scripts to do the bulk upload and EML export functions. These scripts rely on the LTER Network Office (LNO) security framework variables of "primary site" and "user id" to determine identity and authorization for updating and managing. Authorization can be either site or network in scope. Administration web forms allow the site designate to manage all the bibliography entries associated with their site and to add new ones.

From: http://search.lternet.edu/biblio/ - select the "(Manage Site Bibliography)" link. The user is prompted for an LTER network ID password and will then see their site bibliographic records sorted by year descending. From here you can view individual entries, page through the entries, search for specific entries, edit individual entries or add new entries. Edit, and delete functions are provided for each bibliography entry. Entry can either be by single entry form or via a bulk upload process. The bulk upload process allows for manual validate, ingestion, processing and deletion. During the ingestion process the duplicates are identified based on title, author, and year fields. During the processing phase duplicates can either be accepted as updates or deleted – this can be done individually or for the entire upload.

Upload an EndNote Import File - will allow you to upload an endnote import file directly into the database and then get some feedback on the process. We set this up as a push instead of a pull because different sites are usually ready at different times. Note: The EndNote Import file format is based on the Refer / BibIX tag system although it has undergone

some evolution to accommodate new reference types. (See: http://savanna.lternet.edu/reports/endnotetags.php).

Upload an EndNote Import Stream from a URL - will allow you to upload an endnote export data stream directly into the database from a URL. This will also drive the harvesting function which isn't complete yet. Box 1-2 demonstrates the URL provided by Wade Sheldon at the Georgia Coastal Ecosystem LTER (GCE) that can be tweaked to provide the whole bibliography or parts, including only new publications.

Box 1-2 GCE Bibliographic Database URL API Example
http://gce-lter.marsci.uga.edu/lter/asp/db/endnote_export.asp?MinYear=2003&MaxYear=2004&LTERSubmit=no

FUTURE DIRECTIONS

We continue to work on the robustness and usability of the interfaces and the data quality. At this point there are a number of data quality issues that need to be resolved with existing data.

1. Designation of the publication as LTER funded. Some sites track this while others do not. We're trying to make it as easy as possible to designate.
2. Creating individual author entries. EndNote now supports individual author entries and those sites that use them are in good shape. For those sites that don't we are trying to parse the existing author strings into individual authors. Once we're done the site can download the EndNote file to replace there local one. This level of granularity is necessary to be compliant with EML and to allow the linking of individuals to publications in the database.

In addition, we've taken on some new requirements that have been requested to make the all-site bibliography more useful:

1. the implementation of a harvest function,
2. the capability of providing virtual bibliographies for site web sites,
3. a web services interface for updates, and
4. linking individuals in the LTER personnel database to bibliography entries.

Virtual Bibliographies - At press time this feature isn't completed yet but work is in progress. This will allow sites to download the PHP code and implement the search routines on their site web page as if they were local or have them implemented at LNO under site.lternet.edu/biblio.

AVAILABILITY

The php code, perl code, java code, database model and configuration examples will be archived on LNO CVS for this project (http://cvs.lternet.edu).

ACKNOWLEDGMENTS

This article is based on work supported by the U.S. Long Term Ecological Research Network Office which is funded by National Science Foundation Cooperative Agreement No. DEB-02-36154, and the University of New Mexico. T hanks to Troy Maddux and all the LTER Information Managers that contributed data and ideas to this effort, particularly Peter McCartney, Wade Sheldon, John Campbell, and Eda Meléndez-Colom.

REFERENCES CITED

Brunt, James, and Troy Maddux. 2002. LTER All-Site Bibliography 2002 – Update. Databits Fall 2002. (http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/02fall/).

Chinn, Harvey, and Caroline Bledsoe. 1997. Internet access to ecological information the US LTER All-Site Bibliography Project. Bioscience 47(1):50-58. (http://www.aibs.org/bioscience/bioscience-archive/vol47/jan.97.computer.html)

EML Survey Response Summary Report

- Mark Servilla (LNO)

INTRODUCTION

The Fall 2004 LTER Network Information Manager's meeting held in Portland, Oregon provided the venue to solicit responses for an ad-hoc survey on the resources required by each LTER research site to implement a strategy for generating EML. In turn, the LTER Network Office Network Information System group responded to the survey data with a list of five recommended actions:

1. Generate EML based on the current Data Table of Contents (DTOC) database and sufficient to meet Level 1 compliance as defined within the LTER Network EML Best Practices.
2. Generalize and archive the FCE Microsoft Excel template and PERL conversion software for use by other sites.
3. Design and implement an EML Virtual Help website for an open discussion forum and a more static Frequently Asked Questions document.
4. Design and implement a form-based EML editor that compensates for features lacking in Morpho 1.5.
5. Design and implement a workshop (or a series of workshops) that provides specific information on EML tools such as editors, programming concepts (e.g., PERL), XSLT, XML Schema, and EML strategies for RDBMS metadata storage.

Each site was requested to prioritize the list of actions, and to provide any additional comments on the topic. The following report summarizes this effort.

METHOD

Nineteen of the 26 LTER Network research sites responded to the final request for prioritizing; responses varied from brief statements of need, simple prioritized lists, to prioritized lists augmented with additional requests for resources and/or comments regarding the purpose of EML. Ranking was measured from 1 to 5 in order of lowest to highest priority (where the ranking could be defined). Many sites responded with only a single priority action and/or an action that was not part of the original set – in this case, a ranking of 0 was given to each action not evaluated by the site and the highest value given to the single selection. Actions not in the original list were noted outside of the ranking in text. Analysis of ranking was divided into three categories: (1) all responding sites (Figure 5-1), (2) those sites currently generating EML (Figure 5-2), and (3) those sites that are not generating EML at this time (Figure 5-3).

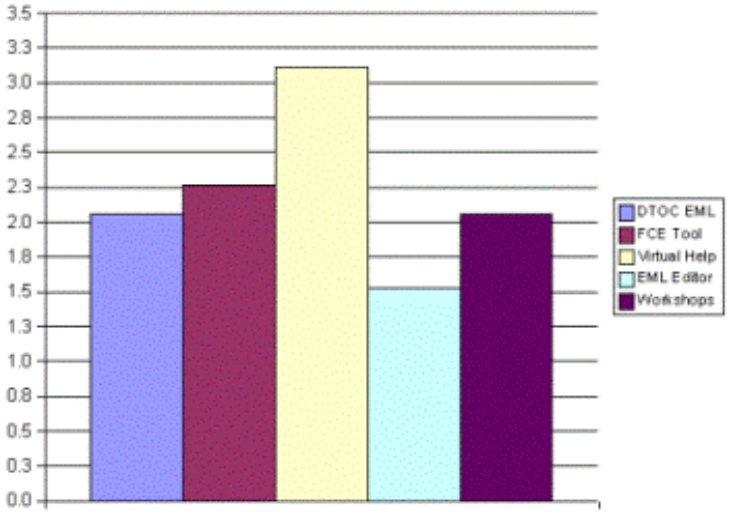

Figure 5-1- Ranking of actions by all respondents.

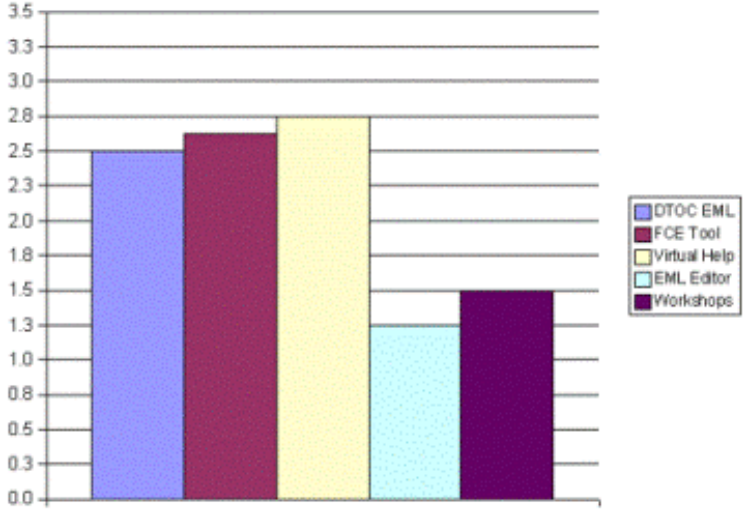

Figure 5-2- Ranking of actions by those site without EML.

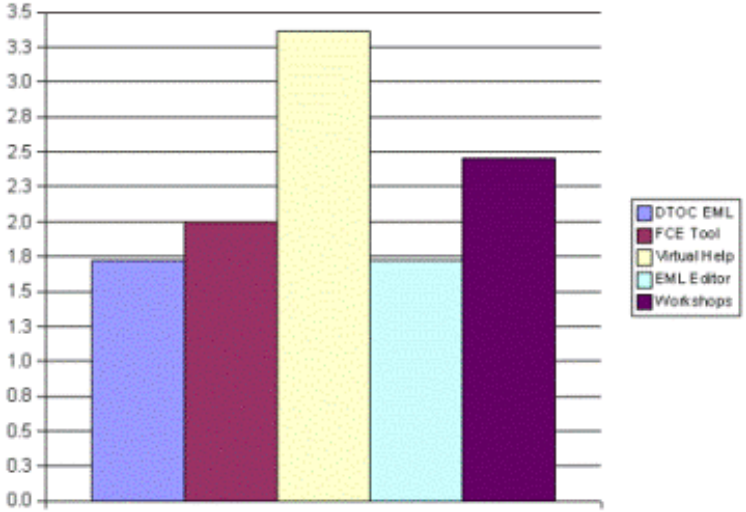

Figure 5-3- Ranking of actions by those sites with EML.

DISCUSSION

In all cases, the highest ranked action was for an EML Virtual Help Desk to be available through the LNO website, while the lowest ranked action was the addition of a new EML Editor. The EML Virtual Help Desk has the potential to benefit all sites by providing short-term and long-term assistance through an FAQ and non-moderated forum. A new EML Editor tool, however, would require considerable resources and time – thus, reducing the benefit to those sites needing immediate assistance with EML, and likely not applicable to those sites already generating EML.

For sites that do not currently generate EML, the second and third highest ranked actions were for standardization of the FCE Excel to EML tool and the conversion of the LTER Data Table of Contents to EML, respectively. Providing EML specific Workshops was ranked only slightly higher than providing a new EML Editor tool. It is assumed that standardizing the FCE Excel to EML tool would provide a short-term solution and a well-known interface for entering metadata and generating EML; one that could be used by site IMs as well as researchers and staff without the direct assistance of the LNO. The conversion of DTOC to EML by the LNO would also provide a method for sites to quickly achieve entry level EML without excess effort. One survey respondent commented that having LNO perform the DTOC to EML conversion may only prolong the absence of more complete EML by providing valid, albeit sparse, content.

The second and third highest ranked actions for those sites that currently generate EML were EML specific workshops and the FCE Excel to EML tool, respectively. The conversion of DTOC to EML was ranked equal to the lowest priority action of a new EML Editor tool. In this case, EML specific workshops appear to have greater appeal when there is no immediate need for an EML generating solution – such workshops may advance the understanding of current tools and/or the general design and use of EML, but may not necessarily provide an EML generating quick-fix.

Many respondents mentioned the need to develop a standardized relational database model for storing metadata; one that would include both a simple entry form and the means to generate EML. Development of such a model by the LNO has the benefits of “economies of scale” for distribution, installation, and maintenance, as well as providing a better metadata management solution for those sites without such a system. Note that the LNO does not currently have a relational database model for storing metadata, and would require time to study and propose a model. At this time, the LNO does not recommend local installation of the KNB Metacat due to the lack of tools that can access and exploit information within the Metacat.

PROPOSED ACTION PLAN AND/OR STATUS

1. The LNO NIS group proposes the immediate development of an EML Virtual Help Desk that includes a Frequently Asked Questions section to be populated with common questions/answers about EML and an online forum that allows questions and answers to be posted without direct moderation. The LNO recognizes that success of an EML Virtual Help Desk depends greatly on input from expertise within the LTER community, and perhaps beyond. The LNO will prepare an executive summary of this effort, and include a preliminary schedule of development and release.
2. The FCE Excel to EML tool has already been reviewed by the LNO NIS group and is currently available through the LTER CVS at http://cvs.lternet.edu. Both FCE and the LNO are updating the tool as necessary to provide needed functionality and bug-fixes. The LNO will also maintain a sampling of completed Excel documents provided by sites as examples for other users.
3. The LNO, in collaboration with KNB, has provided an EML Metadata Workshop oriented to information managers that was held 2-4 February 2005 at the LNO in Albuquerque, New Mexico. Eight LTER Information Managers attended this workshop (refer to http://knb.ecoinformatics.org/knbworkshop2.jsp). The LNO will survey he LTER community to ascertain the need for additional workshops.
4. The LNO is currently reviewing the LTER Data Table of Contents, in conjunction with other site specific metadata, to determine the feasibility of a DTOC to Level 1 EML conversion (refer to LTER EML Best Practices at http://cvs.lternet.edu/cgi-bin/viewcvs.cgi/emlbestpractices/ for a complete description of Level 1 EML). Dr. Inigo San Gil, the LTER/NBII Senior Support Analyst, is in the process of updating all DTOC HTML to XHTML for an initial proof-of-concept conversion of DTOC to EML using XSLT. A brief report of this study will be available in the near future.
5. Due to the low ranking for the development of a new EML Editor tool, the LNO will not address this action in the immediate future. The LNO, however, will evaluate the level of effort to modify the KNB Morpho, an open-source EML editor, to meet the needs of the LTER Network. This evaluation will take place as time permits, but no formal schedule will be outlined at this time.
6. The LNO will immediately begin to evaluate the level of effort required to design and implement a relational database model for metadata management specifically oriented to the LTER community. In addition to the relational database design, this model will include an input interface and an output report generator that will create text, HTML, and EML-based documents.

The LNO will prepare an executive summary of this effort, and include a preliminary schedule of development and release.

KNB Data Management Tools Workshop

- Samantha Romanello, LTER Network Office (SEEK)

The First KNB Data Management Tools Workshop was held September 28-30, 2004 at the National Center for Ecological Analysis and Synthesis (NCEAS), UCSB, Santa Barbara, CA. Interest was so strong that a second workshop was scheduled to reach a larger number of individuals in the data management community - particularly LTER information managers.

The Second Knowledge Network for Biocomplexity (KNB) Data Management Tools Workshop was held February 2-4, 2005 at the new Informatics Training and Software Usability Testing Laboratory. This new, state-of-the-art facility located at the UNM CERIA building is equipped with 21 Dell Optiplex GX280 workstations with dual monitors and an advanced audio-visual system for instruction. The three day intensive workshop was designed to give data managers additional knowledge, skills and tools for managing ecological data at field and research stations across the country. The format included presentations and hands-on, computer-based experiences. Participants were able to upload data and metadata using the newest version of the Morpho metadata software, create and edit Ecological Metadata Language (EML) using XMLSpy, create a login and registration with KNB which allowed users to register their own data and provided access to data stored within Metacat, and create and run a Metacat Harvest and explore the new scientific workflow software Kepler.

LTER was well represented at the workshop by both presenters and participants. Many of the sessions were led by individuals from the LTER Network Office. Mark Servilla (LNO), NIS Lead Scientist, provided a presentation on the LTER EML Best Practices and a hands-on sessions using basic EML via XMLSpy. Duane Costa (LNO), NIS programmer/ Analyst, presented on EML Harvesting and conducted a hands-on session with the new Metacat Harvesting software as seen in the Fall 2004 issue of DataBits. Deana Pennington (SEEK), Research Assistant Professor on the SEEK project presented on Ecoinformatics and Taking metadata (EML) to the next level and a provided a hands-on demonstration session with Kepler: A System for Scientific Workflow also seen in the Fall 2004 issue of DataBits. Samantha Romanello (SEEK), Mellon Post Doctoral Fellow, presented on and provided a hands-on session with the newest version of the Morpho metadata management software. Additionally Laura Downey (SEEK), Senior Usability Engineer on the SEEK project, facilitated group discussion on the usability issues surrounding Morpho.

LTER information managers participants included: John Anderson (JRN), Steve Bauer (CDR), Nicole Czarnomski (AND), Hap Garritt (PIE), James Laundre (ARC), Eda Meléndez (LUQ), Theresa Valentine (AND), and Jonathan Walsh (BES). Other participants included data managers from academic institutions such as Cornell, University of Kansas and University of Hawaii as well as international representation from the Taiwan Forestry Institute, the National Taiwan University and the Departamento de Technologia da Informacao of Brazil.

Both workshops were sponsored by NCEAS, LNO and SEEK. The KNB website (http://knb.ecoinformatics.org) contains a complete workshop description, agenda and slides.

GCE Data Search Engine: A Client-side Application for Metadata-based Data Discovery and Integration

- Wade Sheldon (GCE)

INTRODUCTION

As the number and diversity of data sets in the GCE Data Catalog have grown over the past five years, GCE investigators have found it increasingly tedious to find and download all project data relevant to their particular questions or analyses. Over the past year we have also significantly expanded the scope of the GCE Data Portal web site to include more ancillary near-real-time and historic data sets relevant to the GCE site; however, most portal data sets were not LTER-funded and are therefore not included in the GCE Data Catalog, requiring investigators to find and download data files using an entirely different web interface. Consequently, the information management effort required to help users locate and integrate data for their research projects has been steadily increasing, limiting the resources available for other IM activities. A more comprehensive end-user solution for data discovery, access and integration was clearly needed.

SEARCH STRATEGY

The first challenge was to identify the basic search strategy to use, including the metadata source and content to target. At GCE, we primarily store metadata for core project data in a normalized relational database management system (RDBMS), which supports very comprehensive metadata-based searches using SQL. However, metadata for hundreds of ancillary and provisional data sets are not currently managed in this database and would therefore not be searchable. In order to support searching of all GCE data sets, we chose instead to initially target the structured metadata and data storage standard we developed in the first year of our project and continue to use for primary data processing, distribution and archival (i.e. the GCE Data Structure specification). This standard, based on MATLAB® structure arrays, combines parseable documentation metadata, attribute metadata, data table and QA/QC information into a single computer-readable data package (1,2). The metadata content stored in these data structures is a complete implementation of ESA's FLED standard as described by Michener (3), the same comprehensive standard on which our RDBMS and much of EML is based.

In order to efficiently search for information stored in data structures, which are typically archived as MATLAB binary files on a computer file system, we first developed a comprehensive file indexing application. This application evaluates data structures in any number of directories and subdirectories and generates an optimized search index structure containing complete file details and searchable metadata. We initially chose the following metadata content to index, however the application logic is generic and can be modified to index any available content:

- **General Metadata**
  - Title (text field)
  - Abstract (text field)
  - Keywords (text array)
  - Data Set Themes and Core Areas (text field)
  - Methods (text field)
  - Study Sescriptors (text field)
  - Authors (text field)
  - Taxonomic Names (text array)
- **Temporal Coverage**
  - Study Begin Date (serial date)
  - Study End Date (serial date)
  - Public Release Date (serial date)
- **Spatial Coverage**
  - Study Sites (text array)

    - West Bounding Longitude, decimal degrees (floating-point number)
    - East Bounding Longitude, decimal degrees (floating-point number)
    - South Bounding Latitude, decimal degrees (floating-point number)
    - North Bounding Latitude, decimal degrees (floating-point number)
- **Data Table Attributes**
    - Attribute Names (text array)
    - Attribute Units (text array)
    - Attribute Variable/Semantic Types (text array)

For most GCE data sets, all targeted content can be parsed directly from corresponding metadata fields; however, we quickly realized that spatial and temporal coverage metadata for some classes of data sets are often incomplete. For example, data sets for many hydrographic studies only include study site metadata for the primary cruise transect of interest but not the various marsh-oriented sites also intersected by the cruise track. In addition, data sets produced by investigators for their own use (another target for this technology) often contain detailed temporal and spatial information in the data table but no corresponding coverage metadata at all. Consequently, we also included data mining logic to augment geographic and temporal coverage metadata during indexing (i.e. using attribute descriptors to identify date/time and geographic coordinate data, perform any necessary transformations, and then run detailed temporal and geospatial analyses and GCE geographic database lookups to populate metadata fields).

In conjunction with the data indexing engine, we also developed a flexible search application for querying indices based on parsing delimited lists of criteria (e.g. Keywords = ctd, PAR; DateStart = 01-Jan-2001; DateEnd = 01-Jan-2002; Columns = Salinity; ...) and returning file details and descriptions for all matching data sets. Various text comparison options are supported (i.e. contains, starts with, exact match), but these options are currently set in a configuration file on a per-index-field basis to simplify the query syntax. Negative search criteria can also be specified for any text field or text array field to filter out corresponding matches (e.g. 'Keywords = ctd, -PAR' matches all data sets containing the keyword 'ctd' but excludes any matches that also include 'PAR'). Additional search options can be specified to force case sensitive text comparisons, set bounding box comparison type (fully contained or intersecting), and set overall query type (all criteria matched, any criteria matched). Logic is also included to support compound-field comparisons (e.g. 'Column+Units = Salinity PSU').

In designing these applications, we tried to strike a balance between high specificity (fast development, higher performance, poor reusability) and high generality (slow development, lower performance, higher reusability). We chose metadata content, query syntax and indexing strategies optimal for current GCE needs, but used generalized approaches that could easily be adapted to other tasks in the future. As a result, we were able to achieve very high performance despite MATLAB's relative lack of text processing sophistication. For example, fully indexing 370 complex data sets (i.e. over 300 megabytes of data) takes approximately 90 seconds on a modern PC, and complex searches on this index can be executed in 0.05-0.25 seconds, providing instantaneous results for users.

USER INTERFACE DESIGN

After completing the data indexing and search applications, we developed a comprehensive graphical user interface (GUI) search engine application for building and managing search indices, defining queries, and managing result sets from searches (Figure 7-1). Several different prototype designs were considered, drawing inspiration from the ongoing LTER Metacat Query Interface design process and NBII Mercury search interface, but in the end a multi-paned, single-form design was chosen for the initial implementation to simplify the process.

The top pane includes a scrolling list of all indexed paths and the number of data sets each contains, with buttons for adding and removing paths and refreshing the index to remove deleted files, add new files and re-index modified files in all listed directories. Below the path list, various GUI controls are

also included for entering search criteria and search options to create a query, plus the main 'Search' button for executing the search. Geographic bounding box coordinates can be entered manually, or selected by dragging a box on an interactive map of the GCE study area.

The middle pane contains a scrolling list of all successful queries, along with buttons to manage this list. Prior queries can be reloaded from this list at any time to fill in search criteria fields, allowing users to build up standard queries which they can modify or re-execute against new or updated indices. Query logging can also be disabled and this pane can be hidden to simplify the form and make more room on screen for search results.

The bottom pane contains a cumulative list of all data sets returned from various queries the user has performed. The general location (local or web, see below), accession id, title and study date range are displayed for each data set. Double clicking on any record loads the data set and displays complete formatted metadata in one of several user-selectable styles. The upper button panel in this pane can be used to manage the result set (i.e. sort, select, clear records), and the lower button panel loads the selected data sets into various GCE data analysis applications for editing, visualization (plotting, mapping), statistical analysis, and other operations.

In addition to the main form controls, menu options are also provided to set various user options and to perform batch operations on selected data sets, including copying, exporting in various text and MATLAB formats, and merging to create a composite data set (i.e. by performing a metadata-based union, matching columns by name, data type and units, and merging metadata contents and QA/QC flags). Users can also save complete "workspace" files to disk (i.e. containing the search index, query history, dialog selections and result set), allowing them to persist individual search sessions and reload them for instant start up. A default workspace file is also saved on program exit, so users can pick up exactly where they left off without needing to create or load a new search index.

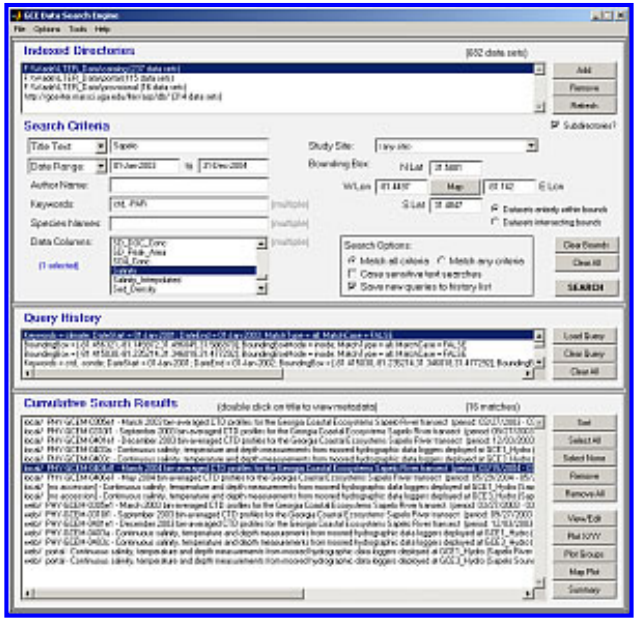

figure 1. GCE Data Search Engine interface
(click on thumbnail to view larger image)

TESTING AND DEVELOPMENT

Input from various GCE investigators and technical staff was solicited throughout the design and development process, and a number of refinements were made based on this feedback. For example, users requested the ability to search for data sets that contained a specified date, so support was added for both date range and contains date queries (i.e. selected using a drop-down menu on the search form). Feedback solicited from GIS experts at UGA was also used to refine the logic for intersecting-type bounding box searches in order to eliminate spurious matches to un-sampled interior regions in multi-site survey data sets (i.e. with large overall bounding boxes). Additionally, early user testing was also instrumental in the debugging process, particularly cross-platform file system issues.

After initial testing was complete and the documentation was finished, the new applications were added to the complete suite of MATLAB-based data processing and analysis programs developed at GCE (i.e. GCE Data Toolbox for MATLAB), and a series of beta versions were posted on the private GCE web site for project member access in Fall 2004. In March 2005, a compiled version of the enhanced toolbox was also released for public access (http://gce-lter.marsci.uga.edu/lter/research/tools/toolbox_download.htm).

INTEGRATION INTO THE GCE INFORMATION SYSTEM

As stated in the introduction, the main objective in developing the GCE Data Search Engine was to enable users to seamlessly search all GCE data holdings and assemble data sets of interest for analyses. However, the file-based nature of the search system we developed requires that all target data sets be available on a local or network-accessible file system

for indexing. In order to accommodate this requirement, automated routines were developed to regularly archive all data sets in the GCE Data Catalog, GCE Data Portal, and provisional monitoring data in Zip format and upload them to the private GCE web site for single-point download. We also regularly create and distribute CDs to GCE members on request, containing all data sets and the latest version of the GCE Data Toolbox software.

Additionally, a more elegant solution was also identified shortly after the first beta release of the software. Recent versions of MATLAB include support for network file access via HTTP and FTP, so a procedure was developed to substitute local file paths in search indices with relevant HTTP URLs for the GCE web site. Code was then added to the GCE Data Search Engine to support web-based files in search indices (including user registration, transparent downloads, and web cache management), and menu options were added to enable users to download and incorporate pre-generated indices of all publicly-accessible GCE data sets from a stable URL on the GCE web site. This capability *dramatically* increased the utility of the application, allowing users to simultaneously search for data stored anywhere on their local file system as well as the GCE web site, and then assemble, transform and analyze these data in one integrated environment. It has also provided GCE users with powerful batch processing capabilities that previously required custom MATLAB scripting to perform.

This software has also proven useful in other areas of the GCE Information System. For example, data indexing routines were used to develop automated software for generating data set summary and detail web pages (based on standard HTML templates) for automatically-harvested data posted on the GCE Data Portal web site. We can now provide detailed, user-friendly web pages for data dissemination (and visualization) with no database or web application server overhead. With this technology, a single computer with MATLAB and Apache could function as a stand-alone data harvesting platform and high-demand data distribution server.

FUTURE PLANS

This software has already proven very useful for a number of data synthesis projects at GCE, and we will continue to refine it to improve functionality. However, we also intend to use the lessons learned for developing improved web-based query interfaces for the GCE Data Catalog, and standardized LTER data query interfaces.

REFERENCES

1. Sheldon, W.M. 2001. A Standard for Creating Dynamic, Self-documenting Tabular Data Sets Using MATLAB. DataBits: An electronic newsletter for Information Managers, Spring 2001 issue. (http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/01spring/)
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. (http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/02fall/)
3. Michener, William K. 2000. Metadata. Pages 92-116 in: Ecological Data - Design, Management and Processing. Michener, William K. and James W. Brunt, eds. Blackwell Science Ltd., Oxford, England.

Designing a Dictionary Process: Site and Community Dictionaries

-Karen Baker (PAL/CCE), Lynn Yarmey (PAL/CCE), Linda Powell (FCE), and Wade Sheldon (GCE)

We are exploring a strategy of unit and attribute dictionary development as a mechanism for moving toward interoperability of site data and cross-site data. We have been considering how to design a dictionary process such that the dictionaries could

be shaped by existing local data procedures and guided by the LTER community standard, the Ecological Metadata language (EML). We find that dictionary building creates a shared understanding of units and attributes outside the local contexts of an individual role, a project, or a site. This work not only informs local data management but also enhances technological accessability of data and ease of data archiving.

Although there are many aspects of units and attributes that remain to be discussed, including what constitutes a dictionary, we suggest that agreement on a shared dictionary template could provide a structure held in common from which to build. A shared dictionary would provide both a boost into the EML learning curve as well as a platform from which to bridge to alternative or related strategies including ontologies.

Although names and their definitions are seemingly mundane and even trivial concepts, this does not mean that the articulation, exchange, and blending of unit and attribute names are simple matters. Names go to the heart of local work practices and of data interoperability. A dual focused approach on local and community dictionaries would permit the LTER IM community to design and prototype a process for creating a community 'living dictionary'.

Building a dictionary takes time as it involves discussions with site participants from technicians to researchers about the dictionary and the information it represents. Time is required for the conversations eliciting unit histories and attribute information. Further, there is an opportunity for learning through comparison with other site entries and for discussing potential site nomenclature changes. In addition, there is a question of whether the dictionary concept, if found generally useful, could create a focus for annual LTER supplements that are occasionally available.

This work evolves from discussions between LTER information managers at PAL, CCE, FCE, and GCE along with files shared by Corinna Gries (CAP) and Margaret O'Brien (SBC). The notions of a shared unit and attribute template for sites and of an LTER community dictionary process seem potential candidates for working group topics for the annual LTER IM meeting. The LTER Information Management Committee is in a unique position to consider designing a review process for proposed entries to a collective dictionary that would provide a mechanism for local names to be considered as candidates for migration to an LTER community dictionary.

Racing With the Typhoon: Storm Strands Scientists on Taiwan Mountain

- Christine Reilley, Peter Arzberger, Tim Kratz, and Fang-Pang Lin (NTL)

Set high in the mountains of northern Taiwan, Yuan Yang Lake (YYL) has been capturing the attention of scientists for more than 60 years. The subtropical lake, nearly untouched by humans, experiences typhoons each year and is surrounded by ancient cypress forest—fertile ground on which limnologists, botanists, and climatologists can conduct long-term studies of its rich environments and ecosystems.

For this reason, scientists from the North Temperate Lakes (NTL) Long-Term Ecological Research project and the University of California San Diego (UCSD) have been traveling thousands of miles during the past year to study YYL. In collaboration with their Taiwanese counterparts at the Academia Sinica Institute of Botany, the Taiwan Forestry Research Institute (TFRI), and the Taiwan National Center for High-Performance Computing (NCHC), they spent the past year constructing the first-of-its-kind global lake monitoring network by establishing wireless connections to sensors in YYL and several lakes in northern Wisconsin. YYL is particularly attractive to limnologists, because a single typhoon can drop more than a meter of precipitation on the 4.5-m–deep lake, causing rapid flushing. This contrasts with the Wisconsin lakes, which have much longer water retention times.

This trip to YYL required one mission: to deposit wireless sensors in YYL that will gather long-term data about dissolved oxygen at various depths and augment existing sensors for barometric pressure, wind speed, and temperature at various

lake depths. The sensors transmit data to databases at NCHC that can be accessed via a web interface from anywhere in the world, allowing scientists to access frequent data about YYL from their desktop. Just a few months earlier, sensors acquired data during a typhoon, recording phenomenon not observed in the temperate lakes of Wisconsin. During this visit, the researchers were looking to continue their collection of exciting data.

The challenge, however, lay in reaching the remote YYL before a typhoon hit I-Lan county, where the lake is located. Storm trackers showed the typhoon was combining with a monsoon and was expected to reach YYL Monday, October 25—the day they were planning to visit.

“It was a race with the typhoon. Considering the kind of damage typhoons cause and that Yuan Yang Lake is in the mountains, the roads leading to the lake may not have been passable if we waited until after the storm,” said team member Tim Kratz, lake ecologist and director of the University of Wisconsin Trout Lake Station, a field site for the NTL Research Station in Wisconsin.

Concerns about the typhoon, however, hardly diminished their desire to venture to YYL. “What’s a little rain?” said Kratz. Colleague Fang-Pang Lin, Grid Group lead at the NCHC, agreed. “We must try, try to get as close to the sensors as we possibly can. If we succeed, all the better.”

Kratz and Lin would be joined by three associates on the expedition: Tim Meinke, aquatic botanist and buoy technician at the Center for Limnology's Trout Lake Station, Dave Balsiger, NTL information management specialist, and Peter Arzberger, director of the UCSD life sciences initiative.

SUNDAY, OCTOBER 24, 2004

The team started the day in the southern Taiwan town of Kenting, where they held a press conference the previous day about a related EcoGrid project on coral reefs. They began their cross-country trek to I-Lan, first catching a plane to Taipei, then driving in the steady rain of the typhoon. After checking into the Hotel Fukun, known for its hot springs, the crew procured microwavable food, wine, and flashlights from the local 7-11 and life vests from Lin’s concerned relatives, who met them near the hotel. Their 90-minute journey took them through winding mountain roads to YYL, their van stalling three times as it passed through deep pools of rising run-off waters. By the time they arrived at YYL, it was 7:00 p.m., and they were ready to set out onto the lake to deposit sensors in the downpour of the typhoon.

Arzberger remained onshore, as their boat could carry no more than four people. Through the blackness and rain, he could only see illuminated flashlights as evidence of his colleagues’ activity on the lake. “Many times I saw the lights drift…much farther than would seem possible. I was worried that they were drifting,” Arzberger said.

The entire procedure lasted more than two hours. As the minutes passed, the temperature dropped, the wind picked up speed, and the waters rose. “It was time to go back to the hotel,” Arzberger said.

After reaching shore, the researchers waded through knee-deep—and, at times, waist-deep—water to return to their van. In the darkness, some wondered if they were on the right path. Kratz, quick to allay any fears, reminded his colleagues of the worst-case scenario: “Hey, you could only get washed into the lake…which is still very calm,” he said.

Fortunately, they were on the right path and reached the van without being swept into the lake. On their drive down the mountain, the crew members—soaking wet and freezing cold—talked about how they were looking forward to taking a dip in the hotel’s hot springs. But when they encountered fallen trees and a landslide blocking the narrow, muddy road, they realized they would not be reaching their hotel any time soon.

Mindful of the precipitous drop-off to one side, the driver, Wen Chung Chang, carefully turned the van around and drove to the field station, which was on the mountain not far from YYL. While en route, YYL park supervisor Chin-Lung Lin called Fang-Pang Lin on his cell phone with good news: a forester living on the mountain offered to let the crew spend the evening at his home. With heat, food, and a cell phone connection, the small cottage made an excellent place to spend the night.

Before settling in at 3 a.m., they uncorked the wine purchased earlier. Although they were marooned on the mountain, they found two major successes to toast: depositing the equipment into the lake and safely escaping YYL.

MONDAY, OCTOBER 25

The crew members awoke at 10 a.m.; for some it was the longest they had slept while in Taiwan, which Arzberger attributed to weak cell phone batteries and the absence of internet connectivity. Upon awakening, they learned of the destruction the typhoon had caused throughout the night. Two major roads leading from the mountain were blocked by fallen trees, mud, and debris; a bulldozer dispatched to clear one of the road blockages had fallen off the side of the road and down a ravine. The driver suffered only minor injuries. The ground beneath another road crumbled in a landslide, leaving just a curb along the hillside for pedestrians to pass in single file. Several other landslides were reported across the county.

It was obvious they would not be able to leave the mountain anytime soon. It also became clear the sensors they had deposited were not functioning. Colleague Hsiu Mei Chou called Tim Meinke on Fang-Pang Lin's cell phone to inform him that tests she ran from her laboratory showed the equipment was not working. The group traveled back to the field station to try to run diagnostics, but the storm had prevented any sustainable connection. The remainder of the day was spent handling other mishaps, with Chang fixing the van's flat tire and borrowing gas for the van from the field station.

Despite their trials, the crew was able to appreciate the day's positive moments. The cell phone connection was strong, so they were able to speak with friends and relatives concerned about their safety. During one conversation they learned The Taipei Times ran an article about Saturday's press conference. Others asked if they had enough food, which was hardly a concern. Chang, who doubled as a gourmet chef, created delicious fare with few ingredients. After a hearty dinner, the crew made their plans for the following day and settled in for the evening.

TUESDAY, OCTOBER 26

The alarm on Kratz’s watch beeped at 5 a.m., and the crew members awoke, determined to resolve their technical problems on the lake and their travel troubles on the road. Fueled by little more than instant coffee, they jumped into their van and set off to YYL, which was still impacted by the effect of the typhoon. While on the lake, this time shrouded in mist instead of darkness, the team discovered a wet serial data interface (SDI) buss was the source of the technical difficulties. After spending two hours on the lake repairing sensors and collecting water samples through the “glug glug” method, the crew began their circuitous trek down the mountain. They hitched a ride on a Mitsubishi flatbed, which took them as far as it could go —to a portion of road completely blocked by a downed tree. They continued their hike on foot, crossing another obstruction of fallen trees. After walking about 3 km, the group was greeted by two motorists, one riding a Sym motor bike and another driving a 1984 Honda Civic. TFRI director Hen-biau King, who had been instrumental in securing accommodations at the forester’s cottage, had called on the motorists for help and arranged the team’s transportation.

Fang-Pang Lin, with life jackets in tow, jumped on the back of the motor bike, while his colleagues rode down the mountain in the Honda. Foggy conditions, along with the Honda’s broken defroster, made for a slow, difficult drive. During the entire ride, the driver steered with his right hand and toweled off the left side of the windshield with his left, while Arzberger, in the passenger’s seat, toweled off the right side.

“The passengers got a bit nervous when the driver answered his cell phone while driving and cleaning. The drop off the

mountain was still quite severe," Kratz said.

The Honda and the Sym brought them to another landslide of mud and tall trees. After skirting this barrier on foot, the group climbed into another van and drove to what would be their final obstacle—the road that collapsed in a landslide. The group walked across a remaining strip of asphalt and emerged to find yet another van from the Fushan Botanical Gardens to take them back to their hotel. From there, they traveled to Academia Sinica, where their colleagues eagerly awaited their return.

Arzberger was not only grateful for those who provided accommodations and transportation, he also acknowledged his colleagues in Taipei and Hsinchu. Whey Fone Tsai, Cheyenne Chen, Julian Yu-Chung Chen, Grace Hong of NCHC, and King acted as a "human lifeline" to the outside world, Arzberger added.

The team effort, Arzberger said, resulted in a successful mission. The data are now flowing from the sensors for all to see at lakemetabolism.org, and the water samples are being analyzed at Academia Sinica.

Said Fang-Pang Lin, "The memories and lessons learned live on. Next time, we will bring some SDI busses, cell phone chargers, and dry clothes."

PHOTOS AND CAPTIONS

Yuan Yang Lake, pictured during milder weather in 2003, is a sub-tropical lake in the mountains of northeastern Taiwan that experiences typhoons each summer.

A buoy with the team's sensors in the middle of Yuan Yang Lake, collecting data about dissolved oxygen, barometric pressure, wind speed, and temperature at various lake depths.

Tim Kratz, Tim Meinke, and Dave Balsiger (left to right) stand in a thicket of cypress trees leading to YuanYang Lake


To leave the mountain, team members had to hitch a ride on a flatbed, ride on the back of a motor bike, and walk along remnants of a road that crumbled in a landslide.

News Bits

A wireless sensor network project for studying lake metabolism at NTL

- Barbara Benson (NTL), Dave Balsiger (NTL)

A wireless sensor network project for studying lake metabolism is underway as a collaboration between the North Temperate Lakes LTER (NTL LTER), the University of California San Diego (UCSD), the San Diego Supercomputer Center (SDSC), the Taiwan Forestry Research Institute (TFRI), the Taiwan National Center for High-Performance Computing (NCHC), and Academia Sinica (AS) in Taiwan. The project website may be found at http://lakemetabolism.org. A recent workshop funded by the Moore Foundation, “Building Capacity and Linking Infrastructure in the Lakes and Coral Reef Scientific Communities” created plans to expand the lake network and establish a coral reef network. The workshop was held March 7-9, 2005 at Scripps Institution of Oceanography, UCSD and was attended by scientists and information technology specialists from 11 countries. A featured article in this issue of Databits relates some recent adventures in Taiwan of scientists working on the lake metabolism collaboration.

Good Reads

Strategies Supporting Heterogeneous Data and Interdisciplinary Collaboration: Towards an Ocean Informatics Environment

- Margaret O'Brien (SBC)

Karen S. Baker, Steven J. Jackson and Jerome R. Wanetick, "Strategies Supporting Heterogeneous Data and Interdisciplinary Collaboration: Towards an Ocean Informatics Environment", in Proceedings of the 38th Hawaii International Conference on System Sciences, Island of Hawaii, Hawaii, January 3-6, 2005.
http://csdl.computer.org/comp/proceedings/hicss/2005/2268/08/22680219b.pdf

Like other earth sciences, oceanography is dominated by interdisciplinary collaborations, in which participants often possess widely varying points of view, research foci, and data management systems. Shifts in the scale of research to global, multi-platform studies, coupled with an increasing interest in diverse partnerships such as those with local stakeholders and educators, have further challenged the current collaborative methods of ocean scientists and their IM systems. In this paper, the authors introduce the "Ocean Informatics Environment (OIE)," a collaboration of researchers, data managers and social scientists involved with several oceanographically oriented projects located at the Scripps Institute of Oceanography. Two of these projects are LTER sites: PAL (Palmer Antarctic LTER), and he newly instantiated CCE (which is closely affiliated with the CalCOFI program, California Cooperative Oceanic Fisheries Investigation). The authors discuss possible strategies to support and improve systems for heterogeneous data management. The goal of the OIE is to design a cooperative environment that is responsive to the varied aspects of its participants' research and data management practices. Their system trades a hierarchical solution for one where diversity and varied needs of collaborators is viewed as an asset to be preserved, rather than an obstacle to be overcome. These challenges are not exclusive to the realm of ocean sciences; all earth and ecological studies are expanding to encompass increasingly heterogeneous data, and so the concepts introduced here are certainly widely applicable.

Revolutionizing Science and Engineering through Cyberinfrastructure

- Karen Baker (PAL/CCE) and Jerry Wanetick (PAL/CCE)

D.E. Atkins, K.K. Droegemeier, S.I. Feldman, H. García-Molina, M.L. Klein, D.G. Messerschmitt, Pmessina, J.P. Ostriker, and M.H. Wright. 2003. Revolutionizing Science and Engineering through Cyberinfrastructure. NSF Blue-Ribbon Advisory Panel on Cyberinfrastructure.
http://www.communitytechnology.org/nsf_ci_report

The Atkins Report represents one in a series of milestones marking NSF supported computer science and technology digital development initiatives. It builds from earlier National Research Council reports focusing on e-Government (2002) and the environmental science grand challenges (2003) as well as the NSF report on supercomputer centers (1995); it emerges alongside reports on digital libraries (2003) and environmental cyberinfrastructure (2003). Bringing forward a recognition of the need for broad interdisciplinary coordination, it highlights the term 'cyberinfrastructure'. Revolution refers to doing things differently, in this case, with new types of scientific and engineering knowledge environments and organizations. We find new names emerging for such work arrangements, from collaboratory and grid to eScience communities. The report discusses not only new environments but also new roles: "The research community needs more broadly trained personnel with blended expertise in disciplinary science or engineering, mathematical and computational modeling, numerical methods, visualization, and sociotechnical understanding of grid or collaboratory organizations." Timely reading, as we consider the next decade of LTER science, the Atkins report opens the door on critical contemporary issues by articulating the need to define as well as to build cyberinfrastructure. It is

initially disconcerting to find the notion of cyberinfrastructure remains fuzzy, but perhaps we are fortunate to be afforded some conceptual space. Rather than adopting a strictly technical approach, we may consider the opportunity, in addition to building cyberinfrastructure, of defining and designing cyberinfrastructure, as a part of a multi-dimensional research process rather than apart from it.

Building the Virtual State: IT and Institutional Change

- Florence Millerand (PAL/CCE)

Jane Fountain. 2001. Building the Virtual State: IT and Institutional Change. Brookings Institution Press.

The book addresses the implications of information technology for institutional change in government. Anchored and illustrated through three case studies, the author points out that the challenges that face the government to build a virtual state are not technological but are largely organizational and institutional. A virtual state -a government organized in terms of virtual agencies, cross-agency and public-private networks - requires more than a technical infrastructure for linking the computers of the government, it demands an institutional infrastructure to support new coordinated practices and procedures between a range of organizational systems. Useful both for information management scholars and for data management practitioners, this book helps to understand the frequent gap between the potential of 'objective' technology (the existing technology product or plan) and the reality of 'enacted' technology (the actual implementation of the planned technology) that results from the introduction of the 'objective' technology in pre-existing social relationships, organizational cultures, and institutional structures.

Calendar

April 13-17, 2005 LTER Student Collaborative Research Symposium in Andrews Experimental Forest, Blue River, OR

June 13-17, 2005 Planning Grant meeting, including all Network Science Working Groups (NSWGs), in Santa Fe, NM

June 14-16, 2005 Network Science Working Groups (LTER Network Planning Grant) Meeting in Santa Fe, NM

June 16-17, 2005 NISAC meeting in Santa Fe, NM

August 4-7, 2005 Annual LTER Information Management Meeting in Montreal, Quebec