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

Editorials

News Bits

Good Reads

Frequently Asked Questions

Calendar

DataBits: An electronic newsletter for Information Managers ----- [Spring 2007] Issue ([http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/07spring/])

Featured in this issue:

Welcome to the Spring 2007 issue of Databits! Twenty-two authors or coauthors submitted articles for this issue, which is a testament to the committment of Information Managers to sharing information. The articles represent the diversity of interests within the LTER IM community and highlight a number of current topics. Most notably, there is a discussion about proposed changes in the organizational structure of the IM committee that would better integrate the GIS working group and the Technology Committee. Additionally in this issue, several articles focus on Ecological Metadata Language, describing recent developments and applications. Lastly, a number of LTER sites will be having their NSF midterm reviews in the coming months. The Baltimore Ecosystem Study was one of the first sites to be reviewed so Jonathan Walsh provided a list of some things to think about as sites prepare for these visits. We hope you find this issue of Databits informative and helpful and we thank all those who contributed articles. Enjoy!

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: John Campbell (HBR), Sabine Grabner (MCR)

Feature Articles

Providing Access and Security to Web Forms using CAPTCHA

- John Porter, VCR

The development of "comment spam" has posed a problem for open access to forms on LTER web pages. Content spammers use automated robots to locate forms on the web and to fill in those forms with advertisements for their dubious (and often offensive) products. One approach is to place all forms behind a password-based firewall. However, this approach is not usable for cases where a form needs to be open to a wider community that will not have pre-existing user names and passwords.

Taking the middle ground between fully open and fully controlled are CAPTCHA-based approaches. CAPTCHA stands for "**C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part" (http://en.wikipedia.org/wiki/Captcha). Alan M. Turing in 1950 proposed a test for distinguishing a computer from a person (MIND [the Journal of the Mind Association], vol. LIX, no. 236, pp. 433-60, 1950). CAPTCHA's implement a greatly modified version of that test to limit access to web forms to people, while excluding web robots.

The most common form of CAPTCHA is a distorted word or set of letters contained in a graphical image. The letters are distorted to make it hard to use optical character recognition programs to distinguish them, while leaving them sufficiently undistorted that humans can still read them. However, there are other types of CAPTCHA, such as mathematical CAPTCHA's that query users on the solution to math problems instead.

Now that we've covered the background, on to the specifics. On the VCR/LTER web site we have a form for scheduling LTER resources that has increasingly been hit by content spammers over the last few months. I wanted to protect that form with a CAPTCHA to eliminate that spam. The first thing I discovered was that there are lots of programs for creating and testing CAPTCHA images, in almost every language (e.g., ASP, Ruby, PERL, JAVA, PHP). I decided to focus on PHP-based programs because my primary focus was on web-based applications and I already have PHP installed. I tried installing and using several packages including:

1. **FreeCap** (http://www.puremango.co.uk/cm_php_captcha_script_113.php) — This package has many attractive features and is among the more secure CAPTCHA's I found. However, it requires an Intel-based computer due to dependencies in the fonts distributed with the system. I couldn't get it to run properly on my big-endian Sun system. It requires that your PHP be compiled with the –with–gd option. However, this is typically not a problem because a version of the GD graphics package is distributed with all recent PHP versions.
2. **CaptchaSecurityImages** (http://www.white-hat-web-design.co.uk/articles/php-captcha.php) — This package has an attractive appearance and is very modular, requiring only a single script and with good instructions for installation. Unfortunately, I could not get this one to run on my Sun UNIX system. In addition to the GD package it also requires the FreeType package. Despite having both of these installed on my system, I could not get PHP to recognize and use the FreeType during the configuration process.
3. **QuickCaptcha** (http://www.web1marketing.com/resources/tools/quickcaptcha/) — At last, one that worked for me! As for all of the above, it requires the GD package that comes with the PHP installation. Instructions are minimal. Primarily you get a web page with an example CAPTCHA interface. I looked at the page source to figure out how it worked and then altered that form to meet my needs. The program requires the installation of several PHP and html files in each directory where you have a form you want to protect, which is less convenient than some of the others. I can also think of a few ways (which I won't mention here for obvious reasons) that might allow it to be circumvented under some circumstances. However, nothing beats something that works!

Check out my implementation of CAPTCHA at: http://www.vcrlter.virginia.edu/abcrc/siteman/local/boatplan/qc_form.html. It uses the CAPTCHA at the front end, rather than as part of the form, since building the CAPTCHA into the form itself requires alteration of the program to which the form data is sent to validate the CAPTCHA.

There are some other solutions that don't even require installing CAPTCHA software on your system. TRYNT (http://www.trynt.com/trynt-captcha-api/) provides a CAPTCHA service where you request the image to be generated on their system rather than yours.

Do you need CAPTCHA's? The big question is whether you are experiencing significant problems with comment spam. A second question, is "how critical is it that visually handicapped individuals access your forms?" Blindness, poor vision and color blindness can all make CAPTCHA's difficult or impossible to read, thus blocking potential users. The final question, "is how secure must the form be?" CAPTCHAs are not unbeatable (some spammers actually hire people in third world countries to post spam), they are merely an impediment to less motivated spammers. If you really need security, a password or certificate-based system may be in your future!

Practical Distributed Computing Approach for Web Enabling Processor-intensive Programs

- Wade Sheldon (GCE)

Introduction

Providing Internet access to processor-intensive programs, such as ecological models and analytical work flows, can present many challenges. If a conventional web application approach is used for hosting the program (e.g. direct access via CGI or indirect access via ASP/PHP scripting) then processor bottlenecks can lead to a denial of service (DOS) condition on the server if too many requests are received over a short period of time. The longer the application takes to complete tasks the more vulnerable the system is to DOS, particularly as users accustomed to immediate feedback press the "Submit" button multiple times, queuing up even more requests. Employing a multi-tiered architecture with the web and application layers residing on different computers can prevent one over-tasked program from locking up the entire web server, but the application server is still subject to process blocking. A different approach is clearly needed to control web access to processor-intensive programs. An ideal solution would be to use a grid computing infrastructure to schedule and execute long-running analyses; however, grid technology is not yet widely accessible and most existing ecological models and analytical programs are not grid-enabled.

I recently encountered this dilemma in my work with the Sapelo Island Microbial Observatory when the lead PI asked me to provide web access to bioinformatics tools that I developed for SIMO researchers. One tool in particular, RDP Agent, executes a complex work flow that includes numerous web and database queries and calls to multiple bioinformatics programs to classify bacteria based on 16S rRNA gene sequence data. Run times for RDP Agent jobs vary from several minutes to several days, depending on the sequence library size and Ribosomal Database Project web site responsiveness, so managing server load was clearly a major issue.

In order to solve this problem, I developed a simple software architecture that included the following components:

1. Web application layer for registering new jobs and providing user access to job status reports and results files
2. Database layer for storing user input (i.e. run-time parameters), managing job status information, and storing links to output files
3. Loosely-coupled application layer for running analyses, using a "check out" and "check in" scheme for controlling job concurrency

This architecture has proven very successful, and we are currently using it to "web enable" two cpu-intensive SIMO programs (RDP Agent and SIMO Library BLASTn). Users receive immediate feedback when submitting jobs, including a report on current server load, and they can check the status of their jobs online or opt for automatic emailing of report files. The major benefit for the system administrator is that maximum cpu load can be controlled on each application server simply by adjusting the size of the application pool (i.e. starting up or shutting down program instances), preventing most DOS conditions. Submissions simply queue up until processes are available to deal with them.

SIMO Tools Architecture

A conventional web application approach was used to develop the user interfaces for the SIMO RDP Agent (fig.1) and SIMO Library BLASTn applications (fig.2). Dynamic HTML form pages were developed using Macromedia DreamWeaver and server-side ASP code. A role-based identity management system was included to prevent system abuse and inappropriate indexing of submission pages by web spiders. New users are automatically granted access for 48 hours to try out the applications, and then access is extended following review by SIMO staff. After logging in, users are able to set various run-time parameters and upload data files for analysis directly from the start page of each application. Web form fields are automatically populated based on the user's last submission, providing personalization of all program settings and options. Links are also provided above each entry form for viewing the status of prior submissions and accessing reports for completed analyses.

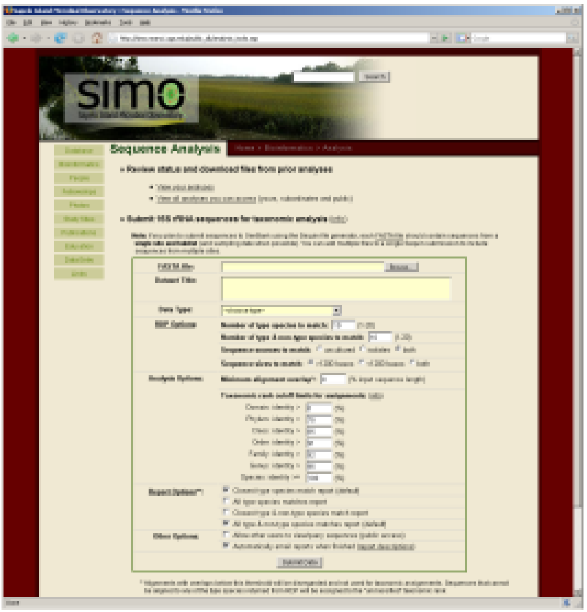

simo
Sequence Analysis
Review status and download files from prior analyses
Submit 16S rRNA sequences for taxonomic analysis

Figure 1. SIMO RDP Agent start screen for submitting data for analysis or viewing prior results
(click on image for larger version)

After submitting new RDP Agent jobs, users are additionally given the opportunity to enter basic research context and environmental metadata for their sequence data set (fig.3). After the analysis is complete this metadata is combined with the classification results to generate fully annotated Fasta-formatted sequence files optimized for batch submission to GenBank using the NCBI Sequin program. This application therefore leverages analytical results to add value to primary data files as well as facilitate submission of environmental metadata to GenBank, enhancing this important community database.

A back-end database was developed using SQL Server to support these applications, with tables for storing user information, submission details, run-time parameters, metadata, analytical results, and report file links. In the case of RDP Agent, an existing SIMO database model was also leveraged for storing sequence data, RDP trees and detailed classification results to support the Sequin output file generation service as well as to build a community database of environmental 16S rRNA sequences that can be mined in the future. Views and stored procedures were developed as a middleware layer, providing abstraction of the database model and a common API for both the web application layer and analytical programs (below).

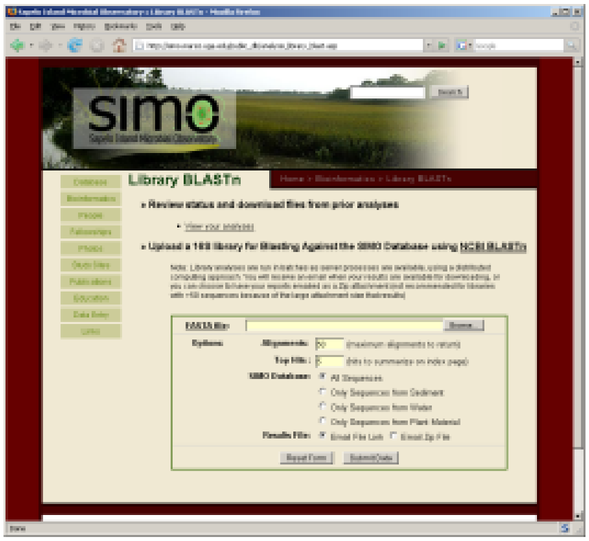

SIMO
Library BLASTn
Review status and download files from prior analyses
View your analyses
Upload a 16S library for Blasting Against the SIMO Database using NCBI BLASTn
FASTA file:
Browse...
Options:
Alignments:
(maximum alignments to return)
Top Hits:
(hits to summarize on index page)
SIMO Database:
All Sequences
Only Sequences from Sediment
Only Sequences from Water
Only Sequences from Plant Material
Results File:
Email File Link
Email Zip File
Reset Form
Submit Data

Figure 2. SIMO Library BLASTn start screen for submitting data for analysis or viewing prior results (click on image for larger version)

Both RDP Agent and SIMO Library BLASTn were originally implemented using MATLAB, so this language was the logical choice for developing the analytical application layer. Key requirements were that program instances support database access via ODBC, timed program execution, network data access via HTTP, and sending email with attachments via SMTP, all of which are supported by MATLAB 6.5 and later (with the optional Database toolbox).

A dispatcher function was developed for each tool to handle communication with the database layer and control program execution. Each time a dispatcher function is called it queries the database to check for any pending analysis jobs. If a pending job is found the dispatcher executes a stored procedure to "check out" the job (i.e. changing the status from "pending" to "running") and retrieve all the run-time parameters. The corresponding application is then called with the specified input, using a try-catch block to trap any unhandled errors that could halt the dispatcher and stop timed execution. After the analysis is completed the dispatcher archives or transmits the output as specified by the contributor, then executes additional stored procedures to update the job status and upload results to the database. If any errors are encountered the dispatcher updates the job status to "failed" and alerts the contributor and system administrator via email. After the analysis is complete, or if no pending jobs are found, the dispatcher exits until called again by the timer.

One key element of this application design is the loose coupling of the analytical program layer and the web application layer. Analyses are only run when a dispatcher instance is available to check out the job, so maximum workload on the server is determined solely by the number of program instances (i.e. MATLAB sessions running dispatcher timers) and not web requests. If a large number of requests are submitted and the program pool becomes saturated, additional jobs simply queue until prior analyses are finished. Another key element is that analytical processes can be run simultaneously on any number of independent systems without concern for duplication. This allows the system administrator to temporarily run additional program instances on other computers to reduce backlogs when many jobs are submitted at once. For example, we usually run four instances of the RDP Agent dispatcher on a dedicated MATLAB application server to handle routine analysis requests, but when new submissions are registered and all instances are engaged the web application sends an email alert and I start up additional instances on my workstation to handle the new requests.

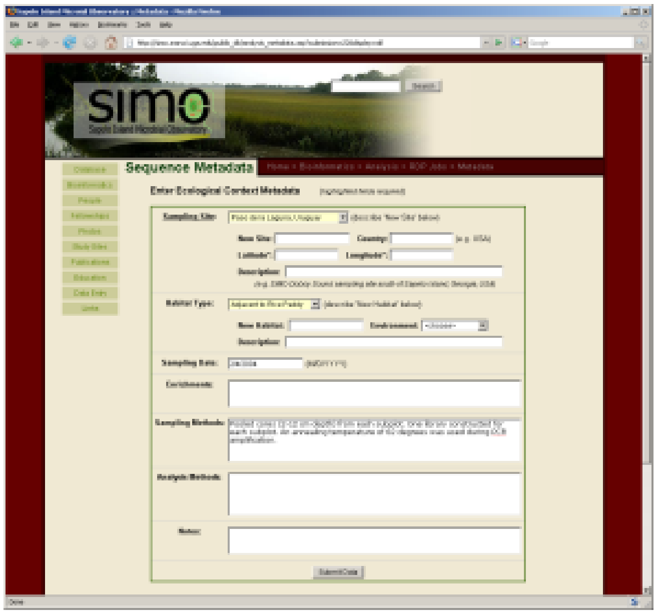


Figure 3. SIMO RDP Agent metadata entry form (click on image for larger version)

Other Potential Applications

I have described one specific application architecture and set of technologies in this article, but I believe this general approach could be applied equally well to many other informatics problems and implemented using a wide variety of alternative technologies. Dynamic web forms are straight forward to develop using popular design tools such as DreamWeaver, and are easy to deploy on various web application server frameworks (e.g. ASP, PHP, JSP, Cold Fusion).

Designing a basic database to store user input and job status is equally straight forward. Implementing the application layer could prove challenging for some target programs, but various high level scripting languages (e.g. Python, Perl, MATLAB) or even shell scripting and crontab can be used for automating program execution. For example, scripts could be developed to parameterize and run an existing model based on web input, then store or transmit the results using a approach similar to the SIMO analysis tools. The only requirement is that the model provide a command-line or programmatic interface so it can be parameterized and called from a dispatcher script.

Conclusion

As the LTER Network moves forward in the decade of synthesis, LTER sites are being challenged to provide broader access to data and analyses on the Internet to support cross-site research. Strategies for enhancing the cyber-infrastructure capacity of LTER site information systems were put forward by the CI-Core team as part of the LTER Planning Grant process, and are currently under discussion by NISAC. Additional CI funding could eventually bring grid computing and other high-end technologies within easy reach of LTER sites. However, the simple distributed computing architecture described in this article can be used to deploy processor-intensive programs like ecological models safely and effectively on the Internet using technology already in place at many sites. In addition to providing broader access to important research tools, I believe this process could also provide Information Managers, web developers and modelers with valuable network computing experience that will ease transition to new technology as distributed computing becomes more central to LTER information system design.

LInks to Additional Information

SIMO web site: http://simo.marsci.uga.edu/
SIMO RDP Agent: http://simo.marsci.uga.edu/public_db/analysis_tools.asp (registration required)
SIMO Library BLASTn: http://simo.marsci.uga.edu/public_db/analysis_library_blast.asp (registration required)

Updating Arctic (ARC) LTER’s Metadata form

- Jim Laundre, ARC

When planning for EML at the Arctic LTER we decided to update our text based form for entering metadata by a form based on an Excel worksheet. Most of our site’s researchers use Excel and there has been some interest in having the metadata and data in the same Excel workbook. After looking at Florida Coastal Everglades (FCE) LTER’s Excel metadata file we developed an Excel worksheet using the following design criteria.

Design Criteria:

- Make it as easy and as clear as possible for researchers to complete the form.
- Follow our previous metadata form where possible but add the necessary fields to bring it to EML Best Practice level 4.
- Use only one worksheet for the metadata form. Thus, it could be easily copied into a researcher’s Excel workbook file with the data.
- The metadata work sheet would not use macros for entering information. This avoids macro security issues.
- An Excel macro using Visual Basic for Applications would be used by the Information Manager (IM) to parse the metadata worksheet to create an EML file. A separate Excel workbook would have the macro, unit information and other common site information needed for the EML file.
- The harvest list would be used for tracking dataset IDs.

The Worksheet for Entering Metadata

The design of this worksheet (Figure 1) is based on the Arctic LTER’s old metadata entry form which was developed in 1989. To ease transition some categories retained their names instead of using the EML tag names. Several new categories were added based on EML Bests Practices but not the entire granularity of EML was implemented. For example Associated Party does not include address information since these are often students, summer RA or others for which addresses change.

Comments are used extensively throughout the sheet to aid in filling out the data. Text boxes are used for sections with more then 256 characters since moving or copying an Excel worksheet truncates the text in cells to 256 characters. Data validation lists are used to created drop down lists for units, measurement scale and number types.

At this time the methods section includes only one text box where all the information about methods are entered. It was decided to be simple and not split out instrumentation, sampling, etc. In a future version we may decide to split out the information.

The variable description (dataTable/entityGroup) section includes columns for attribute name, description, units, measurement scale, number type and missing values. The use of data validation lists for units, measurement scale and number type eases filling in the required values. Not all the EML units are included in the list on the metadata worksheet – only metric ones. Researchers are allowed to enter their own units if none are found on the list. When the file is processed by the IM the units will be checked and flagged if they do not conform to EML standards. The IM will then decide if a custom unit needs to be defined.

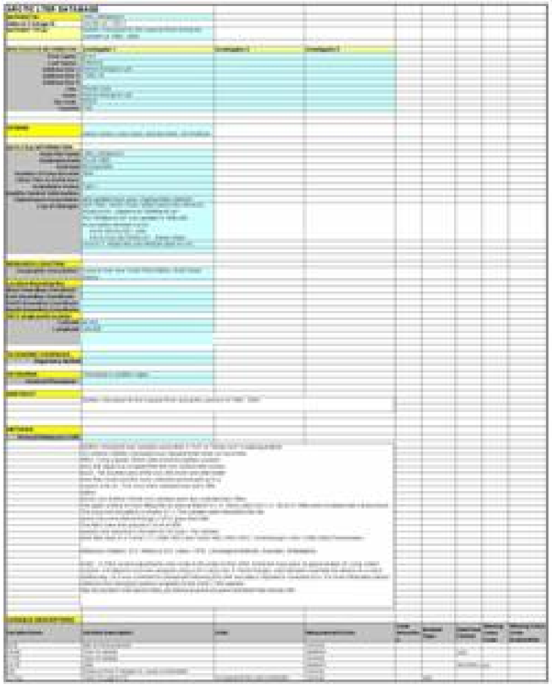


Figure 1. The worksheet for entering metadata
(click on image for larger version)

The Information Manager’s Excel Workbook

This workbook is used by the IM to process the metadata worksheet. It has three worksheets and one macro module:

NameOfRanges worksheet includes the range names. It is used to check for valid names in the metadata sheet.

EML worksheet has the common information along with the EML tags that is included in every EML file. Information such as metadata provider, site intellectual rights, site access, etc. are entered here. In addition there may be information on file locations and URLs. This worksheet is also intended for information that would be LTER site specific thus keeping such information out of the macro code.

Units worksheet is where all the units are listed. This is similar to the FCE Excel metadata unit worksheet. Here custom units and the EML tags are defined along with alias for any of the units. This sheet is used to check the units that are entered by researchers in the metadata form. An alias column is included that allows the IM to enter commonly used alias for EML units.

The VBA macro module, **Excel2EML**, is used by the IM to process the metadata worksheet to create an EML file. The macro’s main procedure steps through the metadata worksheet. Separated procedures are used for checking the metadata worksheet for proper named ranges, for errors in units and for missing required data. The package ID is checked against the harvest list and it will either find the id and check the revision number or assign the next highest number to the dataset. If a new number is used the necessary elements are created and added to the harvest list. The actual EML tags are in separate procedures which are usually functions that return EML elements for the different sections. This helps with error checking and placement of elements in the EML file.

One of the more difficult sections to automate is the URLs for the EML and data files. At present for ARC some assumptions are made about the underlying directory structure where the files are located. However, in a newer version of the metadata form it was decided to have the distribution URL for the metadata and data files entered by the IM in the metadata worksheet. This avoids editing the macro code for changes in file paths or site differences.

Once the metadata worksheet is processed the macro calls HTML tidy.exe to tidy up the code and to get rid of any invalid ASCII characters. Excel often uses special ASCII characters that are not allowed in EML files. It then validates the EML file using the KNB’s EMLparser site on the web.

Future work

There is still more thought and work needed on the metadata entry worksheet and on the macro. Still incomplete is coding for taxonomic coverage and for a more complete methods section. And as in any programming project more error checking and documentation are needed.

Presently Plum Island Ecosystems (PIE) LTER and ARC are using the metadata worksheet. Sevilleta (SEV) LTER and Hubbard Brook Experimental Forest (HBEF) have expressed interest in using the metadata worksheet.

In the future better tools may supersede this simple form. One could also have used Perl to accomplish the same things I have done in a VBA macro. A stylesheet could also have been used.

A Query Interface for EML dataTables

- Margaret O'Brien and Chad Burt (SBC)

1. Introduction

During a time series study, observations are often collected by appending data tables, which quickly become too large to be handled easily by the wide variety of scientists who need them. One common solution is to input the data to a custom relational database and design query interfaces to deliver products. However, once established these systems can be inflexible and problematic to alter as a project's needs evolve. An alternative is to design a more generic application that takes advantage of common practices and can accommodate a variety of data tables. Such an application should depend only on high-quality data and on metadata which follows established community standards. It should not make any assumptions about the metadata content, only its structure.

The recent development of a joint LTER-NCEAS product, the EML Datamanager Library (userGuide.html), has made the process of constructing such an application more realistic. This Java library will be invaluable tool for many applications designed for EML, and we appreciate the opportunity to use it at this early stage. At Santa Barbara Coastal LTER, we are using the library as part of a web application which allows users to query data tables described by EML in Metacat. SBC datasets were used for development, but the application can be applied to EML V.2 data tables which conform to current recommended practices, with a few format restrictions. We plan to deploy the interface through ourdata catalog in the next few months.

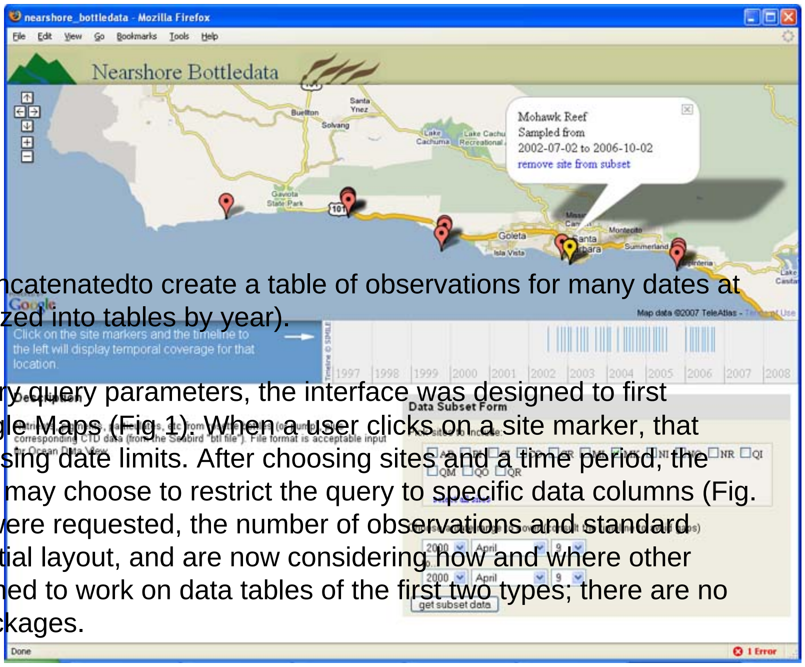
II. Design

Goals: Our primary goal was to facilitate data sharing and reuse among scientists, students and other informed users. Two major considerations drove the design: 1) what queries do our potential users need and how should options be presented to them, and 2) what features of data and EML metadata could we take advantage of? The most common request from potential users was the ability to limit data selection by date and location. Secondly, users wanted to choose the parameters (columns) to be delivered in output. And third, they wanted to create averages (e.g. to generate an average dailytemperature from data that had been sampled at a much higher frequency). Generally, users did not specify merging data tables as highly important.

Data Sources: We identified three basic patterns in data tables which we thought the application should handle:

Type A: the data table includes observations on many dates at many locations.
Type B: the data table includes observations on many dates at a single location, and there may be several packages with identically formatted tables (e.g., a group of weather stations each returning data in the same format in individual tables).
Type C: several data tables with identical format could be concatenatedto create a table of observations for many dates at one or many locations,and then queried (e.g. data are organized into tables by year).

User Interface: Since location and time were given as primary query parameters, the interface was designed to first present the user with a mapof sampling locations using Google Maps (Fig 1). When a user clicks on a site marker, that site's temporal coverage appears on a timeline to aid in choosing date limits. After choosing sites and a time period, the user is presented with a preview of the entire data table, and may choose to restrict the query to specific data columns (Fig. 2). The output is delivered in a zipped CSV file. If averages were requested, the number of observations and standard deviation are included. We are gathering feedback on this initial layout, and are now considering how and where other metadata are best placed. Currently, the application is designed to work on data tables of the first two types; there are no provisions for the concatenation required by Type C data packages.



Mohawk Reef
Sampled from
2002-07-02 to 2006-10-02
remove site from subset

Click on the site markers and the timeline to the left will display temporal coverage for that location.

Figure 1. Screenshot of the initial query page.

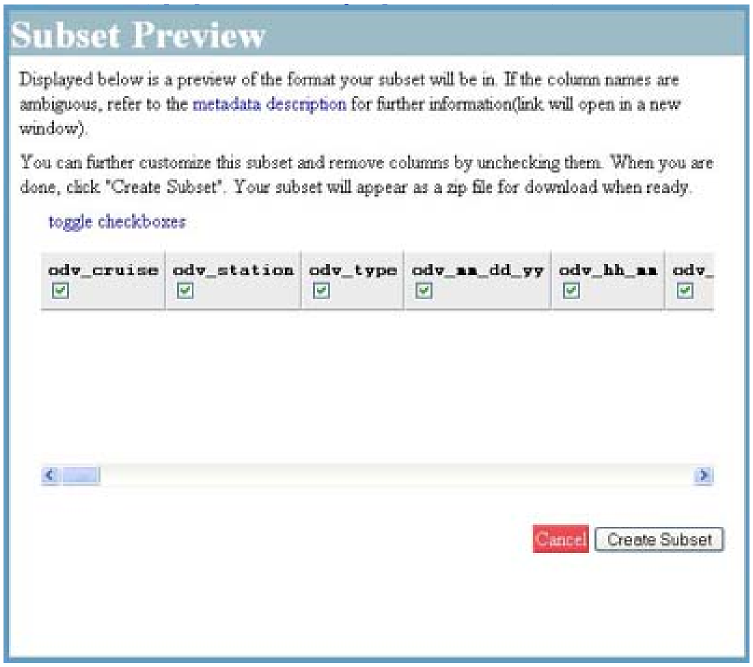
Data and Metadata Requirements: Initially, we tried to limit metadata requirements to only these two:

1. The EML document must have complete and accurate attribute descriptions (i.e., Level 5).

2. The content of the physical/distribution/url/ tag must deliver the data table.

Since sampling locations and temporal coverage were primary query parameters, we explored possible mechanisms for detecting which columns contained this information. The presence of appropriate coverage elements was a reasonable expectation for high-level EML. But due to EML's flexibility, completely automated detection was not practical, and so the following two conditions were added:

3. The table contains a date column, or group of columns from which a date can be constructed (to be specified, see below).



Subset Preview

Displayed below is a preview of the format your subset will be in. If the column names are ambiguous, refer to the metadata description for further information(link will open in a new window).

You can further customize this subset and remove columns by unchecking them. When you are done, click "Create Subset". Your subset will appear as a zip file for download when ready.

toggle checkboxes

odv_cruise	odv_station	odv_type	odv_mm_dd_yy	odv_hh_mm	odv_
☑	☑	☑	☑	☑	☑

Figure 2. Screenshot of the subset preview.
(click on image for larger version)

4. The sampling site(s) are described in geographicCoverage tree(s) at the dataset level. If the table contains multiple sites in a column, then one additional EML structure was required to allow automated detection of the site column (discussed below under "Limits of EML").

III. Architecture

The application is written in Ruby and takes advantage of the built-in Model-View-Controller architecture of Ruby on Rails (Fig. 3). In order access EML, a Ruby client was constructed for Metacat, which will be included in the upcoming v1.7 release. Our application makes use of the EML Datamanager which creates a new table in a local database whose schema is defined by the EML document, and inserts the referenced data into the new table. We used PosgreSQL for the local database. The view layer utilizes the jQuery JavaScript library (http://jquery.com), Google maps, and the SIMILE timeline to create rich AJAX interfaces.

To create a new instance, an operator (probably a site information manager) specifies:

a) The Data's package and table ids. For Type B (groups of tables), a list of ids is specified.
b) A title for the interface page
c) A directory name on the web server
d) The name of the column(s) containing the date element(s). An optional statement is available to enter a PostgreSQL statement to convert a multi-column date into a native Postgres date format.
e) The time intervals on which averaging is to be allowed (options are yearly, monthly, and daily)

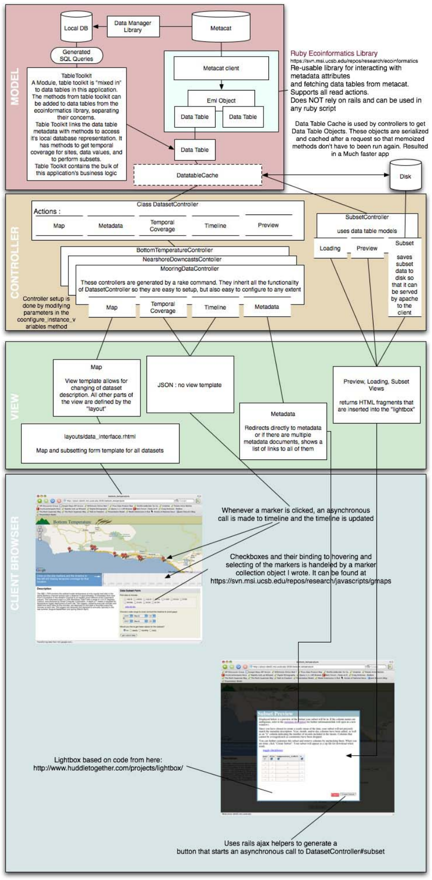

MODEL
Local DB
Data Manager Library
Metacat
Generated SQL Queries
TableToolkit
A Module, table toolkit is "mixed in" to data tables in this application. The methods from table toolkit can be added to data tables from the ecoinformatics library, separating their concerns.
Table Toolkit links the data table metadata with methods to access it's local database representation. It has methods to get temporal coverage for sites, data values, and to perform subsets.
Table Toolkit contains the bulk of this application's business logic
Metacat client
Eml Object
Data Table
Data Table
Data Table
DatatableCache
Ruby Ecoinformatics Library
https://svn.msi.ucsb.edu/repos/research/ecoinformatics
Re-usable library for interacting with metadata attributes
and fetching data tables from metacat.
Supports all read actions.
Does NOT rely on rails and can be used in any ruby script
Data Table Cache is used by controllers to get Data Table Objects. These objects are serialized and cached after a request so that memoized methods don't have to been run again. Resulted in a Much faster app
Disk
CONTROLLER
Class DatasetController
Actions :
Map
Metadata
Temporal Coverage
Timeline
Preview
SubsetController
uses data table models
Loading
Preview
Subset
saves subset data to disk so that it can be served by apache to the client
BottomTemperatureController
NearshoreDowncastsContoller
MooringDataController
These controllers are generated by a rake command. They inherit all the functionality of DatasetController so they are easy to setup, but also easy to configure to any extent
Controller setup is done by modifying parameters in the coonfigure_instance_variables method
Map
Temporal Coverage
Timeline
Metadata
VIEW
Map
View template allows for changing of dataset description. All other parts of the view are defined by the "layout"
JSON : no view template
Metadata
Redirects directly to metadata or if there are multiple metadata documents, shows a list of links to all of them
Preview, Loading, Subset Views
returns HTML fragments that are inserted into the "lightbox"
layouts/data_interface.rhtml
Map and subsetting form template for all datasets
CLIENT BROWSER
Bottom Temperature
Whenever a marker is clicked, an asynchronous call is made to timeline and the timeline is updated
Checkboxes and their binding to hovering and selecting of the markers is handeled by a marker collection object I wrote. It can be found at https://svn.msi.ucsb.edu/repos/research/javascripts/gmaps
Subset Preview
Lightbox based on code from here:
http://www.huddletogether.com/projects/lightbox/
Uses rails ajax helpers to generate a button that starts an asynchronous call to DatasetController#subset

Figure 3. Architecture of the EML dataTable query application. (click on image for larger version)

The application can be configured to distinguish between high frequency and temporally discrete sampling, as well as to control the time frames over which data are allowed to be averaged. The script parses the EML, populates the map, timeline and forms, and then uses the EML Datamanager Library to load the data table into a local database. The process can be automated to recreate the interface when a package has been revised. Loading the table into the SQL database is the time-limiting step, due to the data parsing performed by the manager.

IV. Lessons Learned

Data Quality: The single most important factor limiting the use of a data table in this application was quality, both of data and metadata. Generally, problems with data tables were highlighted by the Datamanager's use of the SQL 'insert' statement, because it validates each row against the metadata description. The data quality issues we encountered were similar to those described by others working on applications for EML documents (e.g., see Data Bits, Fall 2006).

Limits on Scope: Given budget constraints and the fact that this was our first attempt at creating SQL queries from EML, we found it necessary to restrict the scope of this project somewhat:

1. No SQL 'joins'. This application is not designed to merge data tables. It was impractical to anticipate and respond to all possible combinations that might arise. Merging data tables from different packages was viewed as a customized service which can be addressed by future development or on a case-by-case basis. However, the application can represent multiple data tables with one interface. When multiple data tables are selected, the query is applied to each and a zip file containing multiple subset tables is served to the user.
2. Overview of time-series features. A characteristic of regular time-series is that the sampling is predictable, and so it would have been useful for the application to display some additional data features, such as sampling gaps. However, determining the sampling frequency would have necessitated analyzing the data itself, since an element for sampling interval was not available in metadata (discussed below under "Limits of EML").
3. Input: We dropped "Type C" data from the list of data formats to support. Data tables with identical formats and attribute discriptions could be concatenated, but other metadata may not be as easily merged. Repackaging is best left to a separate application.
4. Output: We limited the output of the application to zipped CSV files. Enhancements or other previews (such as plots), alternate formats (such as climDB), or software-specific formats are not supported. Certainly these could be added with further development.

Limits of EML: Within the scope, there were no problems ingesting clean tables and rich, high-quality EML metadata. However, there were two enhancements to EML that would benefit an application such as this:

1. Descriptions of sampling sites: Our users' request that tables be queriable by site meant that our application required a list of sampling locations in the metadata. In practice, projects often use codes as shortcuts for waypoints (lat-lon pairs). A logical structure for storing a list of codes and waypoints is an enumeratedDomain of codeDefinition pairs, in which code=site code and definition=waypoint. Currently in EML2.0.1, <definition> is set to xs:type=”string”, which is insufficient to reliably contain a lat-lon pair. An ideal structure for such a list is codeDefinition pairs in which each definition references the id attribute of a geographicCoverage tree. In fact, this feature has already been requested. In anticipation of this eventual resolution, we adopted the following practice:

a. The site attribute was coded as an enumeratedDomain.
b. The site code is also the id value of a geographicCoverage tree at the dataset level.
c. The geographicCoverage tree was copied to the site attribute.

To detect the correct column, the program logic defines the site column as the entity attribute that contains an enumeratedDomain *and* a geographicCoverage tree, and populates the map from the values in the local tree. If/when referencing has been enabled for definitions in enumeratedDomain, the logic can be edited to detect the entity attribute with an enumeratedDomain whose definitions are references to dataset-level geographicCoverage trees, resolving the reference to retrieve content.

2. Descriptions of time-series: One feature of a regular time-series is that observations occur according to a known sampling interval, but there is no element in EML which defines that interval. Had this value been available in metadata, the timeline could have been formatted so that sampling gaps were evident to the user before constructing the query, instead of displaying only a continuous bar between begin-end dates.

Database Storage Model Considerations: XML and Relational Database Approaches

- James Conners (PAL & CCE)

When writing a computer software application, be it for standalone purposes, web applications, etc. there is almost always the issue of persistent data storage models to consider. This may involve a range of factors such as account settings, application configuration parameters, and scientific dataset characteristics. The question of whether to store this data in XML files or a relational database often arises. The decision requires an assessment of the requirements for each application, taking into consideration the programmatic aspects of development, maintenance, documentation, as well as the inherent characteristics of the data. Searching through developer forums will certainly bring up a variety of discussions on the more technical aspects of the comparison of these approaches by technologists with detailed experience; this article focuses simply on a few higher-level considerations concerning portability, data access, and inherent data characteristics.

In consideration of portability, the desired degree of this attribute will often dictate which is the best storage technology to implement. The greater the focus is on an application that can be used by a variety of users and on different platforms, the more data storage options tend to point towards portable documents. XML documents, simply being text files, are at the very high end of portability. Any technology that can parse text files can implement a storage model that relies on XML. Implementing relational database technology requires, on the other hand, an installation of a particular DBMS (Database management system) such as MySQL, Oracle, etc. on either local machines or a dedicated server that can be accessed by remote instances of the program on client machines. Of course, the last DBMS approach removes a degree of portability with the requirement for connectivity to a remote system. Another aspect of portability deals with the transport of data. XML is perhaps best known for its advantages in transport, that is, in the sharing of data. The XML standards established by the WWWC (World Wide Web Consortium: http://www.w3.org/TR/2000/REC-xml-20001006#dt-doctype) provide for a means of validating XML documents against established DTD's (Document-type Declarations). The standardized schema definition language XML Schema (W3C) builds on DTD’s and provides for a more robust document definition. Either method provides a mechanism for maintaining data integrity during transportation.

Data access is another issue to consider when deciding between XML and a relational database. RDB solutions are known for their advantages in terms of data retrieval. If the stored data must be queried by varying parameters, and the dataset is large, the relational database system has advantages. Relational theory is a well-established field of study and as a result the current DBMS’s are highly optimized for searching through large amounts of data based on parameter specifications. The majority of modern, higher-level programming languages provide interfaces to the well-established DBMS’s thus enabling applications access to the efficiency of SQL and relational structures. With the standardization of the query language like XML Query (http://www.w3.org/XML/Query/) the capability of searching the relational structure of data stored in XML format is being improved. Yet, as the size of the files grows, XML’s serial hierarchical structure and text-based format means that searching through XML documents can significantly tax the performance of an application.

Another deciding factor in choosing between XML and a relational database is whether the data structure at hand is consistent or not. Relational databases are a natural fit for tabular data with a regular, well-defined structure. When the structural properties of the data vary, however, XML might be a better fit. "Because XML syntax records only what is present, not everything that might be present" it can be better suited to records that can vary drastically with regard to missing values or even completely different field sets 1. Ronald Bourret gives an example during a discussion on storing XML files in which he points to medical records where different patients may have completely disparate records with mutually exclusive properties. Accommodating all the possible parameters in a relational database would lead to sparse tables; using XML's explicit structure of available data appears to be a more suitable solution.

In exploring solutions, it is possible to consider arrangements that combine the best features of both XML and relational databases. A common perception of the two technologies is that XML is best used as a document-sharing format, that is, a transport mechanism. On the other hand, a relational database is recognized for its ease of implementing the many aspects of data maintenance and search optimization. Thus, many DBMS's either provide native capabilities for manipulating XML-formatted documents or at minimum provide plug-in functionality for this purpose. The combination of using a database for data storage together with XML documents for transporting of these data thus represents a method well supported by programming languages. One example of combining these technologies is with Ajax web-applications. Storing persistent data in a database backend makes easy work of updating, deleting, and inserting data, and using XML to transport the data utilizes one of HTTP's (Hypertext Transport Protocol) supported mime types and allows the document to be manipulated as an object within local javascript code. With this arrangement, each technology is used where it best supports application development.

A specific local project provides a case study of this approach. The project consisted of building a glossary application for web access to commonly used acronyms. Beginning with an XML document to store these simple term/definition pairs seemed appropriate because the storage requirement was so small. After an editing feature was implemented, however, the need to allow multiple updating transactions prompted the use of a RDBMS. No aspects of the relational storage model are used in the single-table database. Rather, the choice was guided by the concurrent transaction support a RDBMS provides, allowing multiple users to edit the glossary through the web interface.

The decision of whether to implement an XML or relational database storage model is typically case specific. For the storage of persistent data, the particular application of that data or the data itself will offer clues to the best approach for storage. The three issues/factors discussed above–portability, data access, and data characteristics–are far from comprehensive but do represent some important broad assessment categories. Personal preferences and experience will always be strongly influential, along with frequently used conventions. Within some application arenas, there may be an existing, generally more acceptable, approach to data storage. With both popular conventions and traditions used in our local informatics environment comes the benefit of standardization and developer support. And so, as expected, there is no simple or single answer to the question of data storage models.

[1] C. M. Sperberg-Mcqueen, World Wide Web Consortium XML and Semi-structured Data(http://www.acmqueue.org/modules.php?name=Content&pa=showpage&pid=339&page=1)

Information Infrastructure: Transitioning Directory Services

- Karen Baker, Jerry Wanetick, Nate Huffnagle, and Mason Kortz (CCE & PAL)

1. Introduction

What is 'Directory Services' and how can we benefit from this technology? Directory Services (DS) addresses 'scaling' up in terms of resources and users, an issue that arises, for example, when an organization grows from a small, dedicated project or laboratory data management group to a larger more complex and interconnected environment. By incorporating Directory Services into the network, we address the need to use and/or to manage an increasing number of users, servers and application services. DS enables the federation of user and application authentication and authorization information over a distributed network of desktop computers and servers. Individuals carrying out research or providing system support for cross-project and cross-program activities often find the type of information system infrastructure provided by DS simplifies the work arena.

Directory Services have been around for a number of years as part of the ISO X.500 series of networking standards. The most popular and ubiquitous of these has been Sun's Network Information Service (NIS). Electronic directory services include Directory Access Protocol (DAP) and Lightweight Directory Access Protocol (LDAP). The LDAP standard is highly customizable. However, this flexibility introduces inconsistencies when manufacturers modify standard configurations to accommodate specialized product functionality, thus building in cross-platform incompatibilities.

The PAL and CCE sites work within the Ocean Informatics (OI) environment. OI is transitioning from NIS to LDAP as part of an ongoing effort to improve the information infrastructure, in this case with a more contemporary Directory Services approach. Based on consideration of the state of industry standards, the transition is centered on the use of Apple's implementation of Directory Services, known as Open Directory (OD) [http://www.apple.com/server/macosx/opendirectory.html]. This approach was selected because it is based entirely on Open Source projects. This contrasts with other approaches such as Microsoft's Active Directory and Sun's Java System Directory.

Open Directory brings together OpenLDAP [http://www.openldap.org], MIT Kerberos [http://web.mit.edu/kerberos], and Simple Authentication and Security Layer (SASL) developed at Carnegie Mellon University [http://tools.ietf.org/html/rfc4422; http://asg.web.cmu.edu/sasl] into a coherent and secure Directory Services framework. Here Kerberos, along with the SASL-based password server, provides a secure authentication method so passwords are never sent across the network. OpenLDAP provides authorization information, enabling central administration for provisioning infrastructure assets. The OD provides these authentication and authorization services across a heterogeneous network of computer architectures, such as: Sun Solaris, Redhat Linux, Microsoft Windows XP (and now, Vista) and Apple MacOSX operating systems.

Packaging services in a more modular, streamlined fashion allows for certain economies of scale. For instance, when a user is authenticated by the Open Directory authentication services, the user is handed a Kerberos ticket (i.e., certificate or credential) that is then used to allow access to the assets that the user is authorized to use. From the user perspective, what was previously a set of machine specific user accounts with separate passwords and account information has been replaced by a "single-sign-on" environment. From a systems administration perspective, asset management is made easier by providing a central means of handling user account and application information. In addition, the collection of services, as well as who has access to these services, are available in the newly centralized schema. Also, by centralizing user authentication and authorization information, the stage is set to implement tokenized or one-time password technologies to allow secure remote access to local infrastructure.

2. Our Local Experience

In a homogenous environment, implementation of native directory services is fairly simple. If the groups associated with Ocean Informatics were completely Mac-based, bringing all of the computers and services under an OD managed environment would be straightforward. However, this is not the case. The PAL/CCE/IOD infrastructure is comprised of Sun Solaris, Redhat Linux, Microsoft Windows XP (and now, Vista) and Apple MacOSX operating systems. The greatest challenge in implementing OD in this environment has been integrating these disparate operating systems into the OD infrastructure. Each platform - Apple, Microsoft, Sun and Redhat - uses a slightly different approach to directory services, though all start from the same standard and are loosely based on similar technologies. By modifying the default Open Directory schema in MacOSX, we are able to bridge the differences between these platforms and to increase interoperability across diverse, heterogeneous networks.

Incorporating these new technologies into a heterogeneous network environment has taken time, that is, it has not been an effortless process. As a bridge between pre-existing or legacy directory services and the new OD framework, we have adopted a long-term approach with a strategy of phased-in implementation to minimize the impact of change on our user base, and to allow for testing as new elements are enacted. The first step, after populating the directory server with user and server information, was to integrate core services such as email, file sharing and user authentication. Subsequently the division's mail server, file sharing/home directory servers, storage servers were integrated into the OD domain. The next step is to integrate client workstations and new servers into the OD domain. For example: when called upon to repair or install new clients or servers, such as when the IOD business office purchased a new server in Fall 2005, these machines are integrated into the OD domain. In upcoming months, the department ftp, web and collaboration servers, and the other independent servers will be integrated into the Open Directory Services. As of January 2007, 80% of legacy servers, 100% of new servers, 40% of existing clients and 40% of new clients have been migrated.

Services currently supported by OD in our local Ocean Informatics environment include email, file sharing (via AFP, SMB, and NFS protocols), print, website authentication (via Apache, mod_auth_apple, and mod_auth_ldap), and user login services. Directory Services minimizes the need for users to track multiple accounts, and also minimizes user account management overhead, thus allowing users to move from machine to machine without the need for an administrator to create a local account and transfer file and settings on each machine.

The pace of change can be measured in phases over a period of years:

Phase 1 (2004) Started planning in summer; mapping NIS schema to LDAP schema
Phase 2 (2004) Synchronize UID/GID space across standalone servers
Phase 3 (2005) Set up an OD server for testing and pick test clients
Phase 4 (2006) Migrate NIS to OD on a production server
Phase 5 (2006-2007) bring servers and clients under LDAP umbrella

3. Conclusion

Implementing Directory Services is part of Ocean Informatics' ongoing quest for balancing of resource access expansion and sustainable practices. With Open Directory in place, system administrators are able to spend more time attending to the needs of a growing user base and implementing new tools, and less time involved in simple but time-consuming user management tasks, and users only have to track a single account rather than multiple accounts on numerous computational resources. In addition to central user management, OD allows for more refined and centralized control over system resources. One central server can now control disk shares, server connectivity and access to workstations.

Apple's Open Directory fits well with the Ocean Informatics open architecture approach with an emphasis on Open Source Software. Implementing Directory Services in a homogeneous environment is fairly straightforward. The continuing challenge for PAL and CCE within the Ocean Informatics approach, is to implement services in a heterogeneous computational arena with a variety (Mac, Microsoft, Unix, and Linux) of new and legacy systems. The challenge has been to make the interoperability between various platforms as seamless and coherent as possible. Our goal is to increase the availability of information that is platform independent. By using a scalable, network standard for directory services, we can connect multiple machine architectures and technologies in a unifying manner that increases availability and accessibility of digital resources.

Generating EML from a Relational Database Management System (RDBMS)

- Zhiqiang Yang and Don Henshaw (AND)

The Ecological Metadata Language (EML) is an open metadata specification and provides a standard syntax (XML) for LTER metadata (Jones 2003). Implementation of the LTER Metacat, a network-wide data set catalog, demands the creation of EML documents for all LTER data (Costa 2004). Future data integration activities, such as the development of the Trends database, will also rely on complete EML documents (Servilla et al. 2006). Development of EML documents is a data specific process depending on the existing format and structure of metadata, and there are different approaches that can be used to achieve this goal. One common means for storing metadata is in structured relational tables in a relational database management system (RDBMS).

In this short document, a generalized solution for generating EML from an RDBMS is presented and based on architecture originally developed at the CAP LTER (CES 2002). Storing metadata in an RDBMS has the advantage of being able to select and output metadata in various formats including EML. The approach presented here is two-step process of generating EML from the RDBMS: 1) generate a native XML document for metadata stored in RDBMS; 2) transform native XML into the EML document (Figure 1).

Generating native XML

Metadata stored in a site RDBMS are typically organized as attributes in relational tables without any direct correspondence in structure or naming convention with matching EML elements.

Generating EML directly from these stored metadata is challenging and cumbersome, and does require a mapping of local site attributes to the corresponding element in EML. Here, a native XML document is suggested as a connection point between the RDBMS and EML. For this discussion, native XML is an XML document which fully encapsulates all the metadata contained in the RDBMS for a specific dataset. Native XML does not have a predefined schema; the schema of native XML is dependent on the database schema for metadata, but creation and use of a configuration file allows specification of RDBMS tables for inclusion into the native XML.

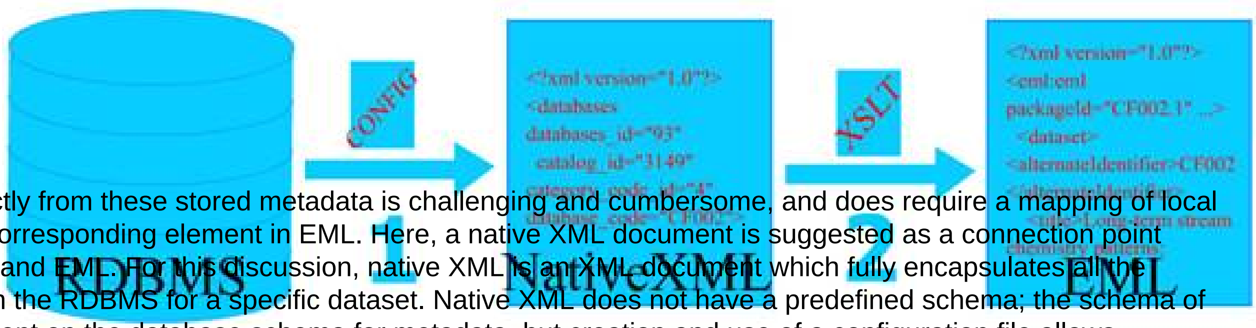


Figure 1. Data flow of metadata from the RDBMS to EML, e.g., from the local RDBMS table *databases* to the EML element: dataset.

Most RDBMS's maintain a data dictionary, which is a set of metadata that contains definitions and representations of data elements stored in the database. For example, SQL Server has two sources to view the data dictionary: the various system tables and the INFORMATION_SCHEMA views. Users can query these two sources directly or use system stored procedures to retrieve information about the database metadata. In the following examples, the SQL Server system stored procedures, sp_columns, sp_pkeys, and sp_fkeys, are relied upon to programmatically access table column (attribute) information, and table primary and foreign keys.

Native XML generation is highly dependent on the data dictionary. To generate native XML an entry point is needed, which usually is a table representing objects corresponding to the EML dataset element. Typically, this starting point is a table with general dataset catalog attributes such as dataset code, title, abstract, and other dataset level metadata with relational links to other tables such as personnel, keywords, or data set entities. With a given starting point, all information stored in the RDBMS related to the given dataset can be retrieved using a standard SQL statement. This process uses extensively the information in the data dictionary, and native XML is generated by retrieving all the columns in the database related to the dataset. For example: suppose there are three tables as shown in Figure 2.

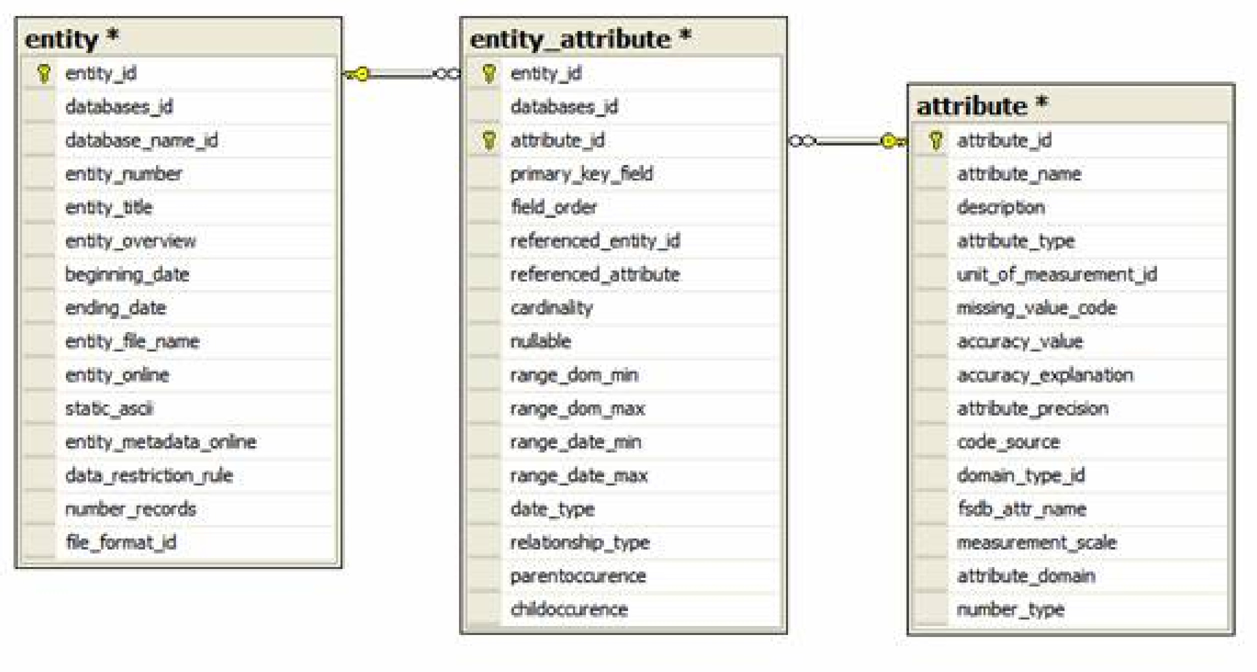
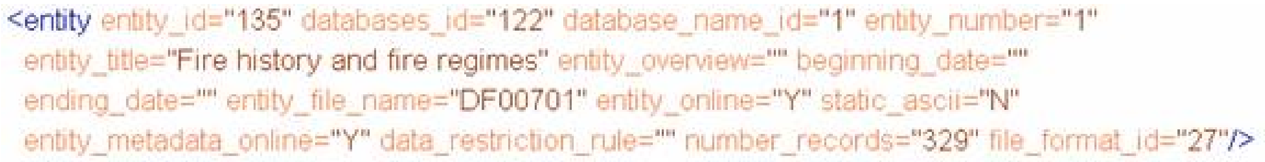


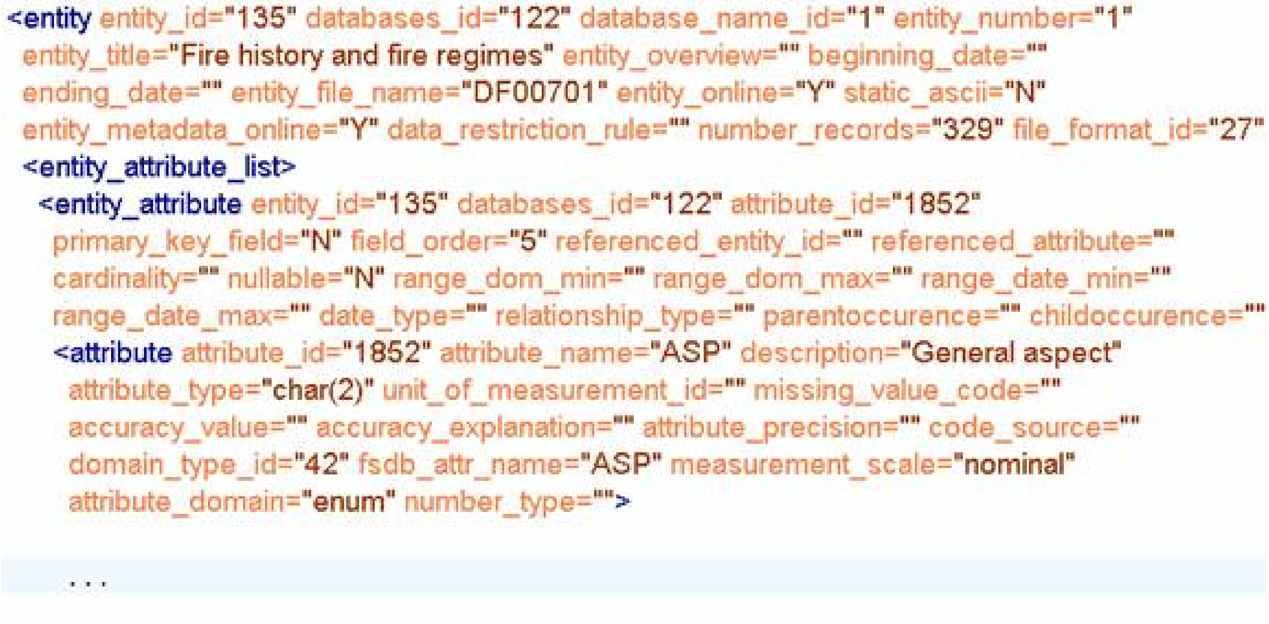
Figure 2. An example of a partial schema (from the Andrews Forest metadata schema). The *entity* table might be directly related to a parent dataset catalog table (not shown).

The native XML for an entity with entity_id = 135 would include the content from a single record in the entity table:



```
<entity entity_id="135" databases_id="122" database_name_id="1" entity_number="1"
  entity_title="Fire history and fire regimes" entity_overview="" beginning_date=""
  ending_date="" entity_file_name="DF00701" entity_online="Y" static_ascii="N"
  entity_metadata_online="Y" data_restriction_rule="" number_records="329" file_format_id="27"/>
```

Since the *entity* table is related to the *attribute* table through the *entity_attribute* pivot table, the native XML generation process will use the referential integrity information in the data dictionary to iterate through all of the records in both *entity_attribute* and *attribute* tables, resulting in a native XML similar to the example in Figure 3.



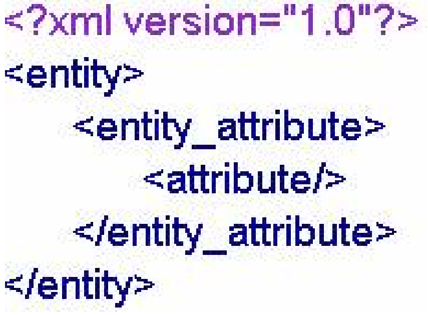
```
<entity entity_id="135" databases_id="122" database_name_id="1" entity_number="1"
  entity_title="Fire history and fire regimes" entity_overview="" beginning_date=""
  ending_date="" entity_file_name="DF00701" entity_online="Y" static_ascii="N"
  entity_metadata_online="Y" data_restriction_rule="" number_records="329" file_format_id="27"
  <entity_attribute_list>
    <entity_attribute entity_id="135" databases_id="122" attribute_id="1852"
      primary_key_field="N" field_order="5" referenced_entity_id="" referenced_attribute=""
      cardinality="" nullable="N" range_dom_min="" range_dom_max="" range_date_min=""
      range_date_max="" date_type="" relationship_type="" parentoccurence="" childoccurence=""
      <attribute attribute_id="1852" attribute_name="ASP" description="General aspect"
        attribute_type="char(2)" unit_of_measurement_id="" missing_value_code=""
        accuracy_value="" accuracy_explanation="" attribute_precision="" code_source=""
        domain_type_id="42" fsdb_attr_name="ASP" measurement_scale="nominal"
        attribute_domain="enum" number_type="">

        . . .
```

Figure 3. Native XML generated for entity_id=135 from the *entity* table, and using the *entity_attribute* table to link to the *attribute* table to list all of the attributes for that entity (entity to attribute is a many-to-many relationship, but only one attribute is shown)

This iterative process of metadata retrieval is continued until all the related information for a given entry point has been exhausted. That is, native XML from all related tables stemming from the original starting point is included.

However, problems frequently occur while programmatically iterating through the related tables, as a primary key may loop back to a previously included table. In the example, *entity_attributes* will refer back to the *entity* table creating a loop between tables *entity* and *entity_attributes*. Program code is written to prevent this loop from occurring by taking advantage of a configuration setting or file. The configuration file, a simple XML file, is established based on the data dictionary and helps guide the process of native XML generation. The configuration file is used to prevent the duplication of table information, describes the hierarchy of related tables, and lists the tables for which native XML will be generated. The configuration file for the example is shown in Figure 4.

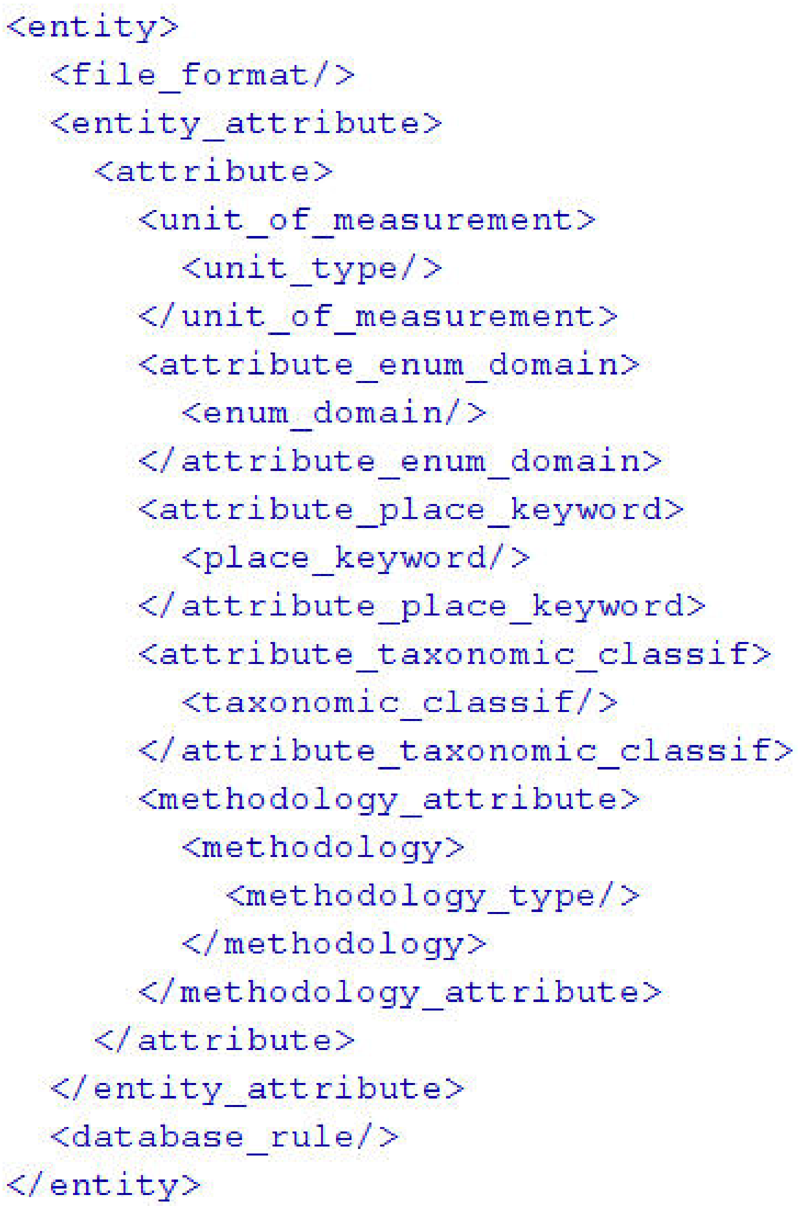


```
<?xml version="1.0"?>
<entity>
    <entity_attribute>
        <attribute/>
    </entity_attribute>
</entity>
```

Figure 4. This example is a configuration file which

specifies the tables for which native XML is to be generated.

A more complete configuration file example is included as Figure 5 and hierarchically represents many other tables associated with the attribute table from the Andrews metadata schema. This example includes unit of measurement, enumerated code domains, place domains, taxonomic classification domains and attribute-specific methodology.



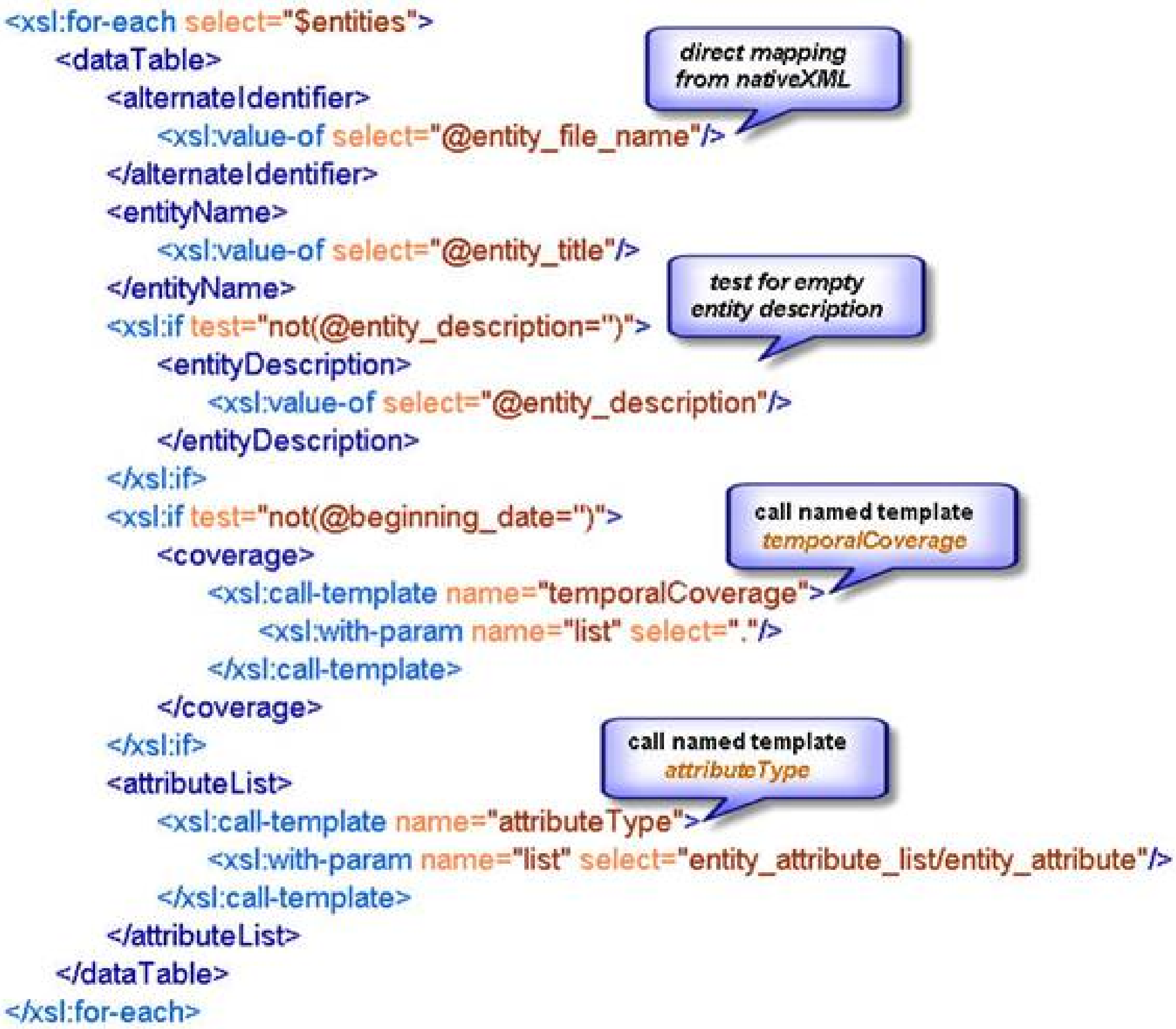
```
<entity>
  <file_format/>
  <entity_attribute>
    <attribute>
      <unit_of_measurement>
        <unit_type/>
      </unit_of_measurement>
      <attribute_enum_domain>
        <enum_domain/>
      </attribute_enum_domain>
      <attribute_place_keyword>
        <place_keyword/>
      </attribute_place_keyword>
      <attribute_taxonomic_classif>
        <taxonomic_classif/>
      </attribute_taxonomic_classif>
      <methodology_attribute>
        <methodology>
          <methodology_type/>
        </methodology>
      </methodology_attribute>
    </attribute>
  </entity_attribute>
  <database_rule/>
</entity>
```

Figure 5. A more extensive example of a configuration file based on the RDBMS data dictionary that includes the tables desired for inclusion in the native XML.

In summary, the configuration file is used to guide the process of creating native XML in conjunction with program code and referential integrity information provided by the data dictionary. The configuration file determines which tables are to be included in native XML generation allowing the exclusion of tables defined within the RDBMS schema.

Generating EML

XSL Transformation (XSLT) is used to convert native XML to a valid EML document. The key aspect of this process is to properly map the local metadata stored in the RDBMS to the corresponding EML <dataTable> elements. The mapping of a database schema to EML elements requires understanding of both the local RDBMS and the EML schema. The mapping process can become complex in the common situation where the metadata database is not directly designed to accommodate EML metadata elements. And while the authors do not attempt to describe the XSLT language, we do provide a short example XSLT stylesheet to help illustrate this process of generating the EML <dataTable> element in Figure 6. For the example given above, the local *entity* table corresponds to <dataTable> in EML. The stylesheet illustrates direct mapping of a native XML element into EML, checks for an empty field before mapping entity_description, and uses named templates as functions or subroutines to map the native XML into EML coverage and attributeList elements.



```
<xsl:for-each select="$entities">
    <dataTable>
        <alternateIdentifier>
            <xsl:value-of select="@entity_file_name"/>
        </alternateIdentifier>
        <entityName>
            <xsl:value-of select="@entity_title"/>
        </entityName>
        <xsl:if test="not(@entity_description='')">
            <entityDescription>
                <xsl:value-of select="@entity_description"/>
            </entityDescription>
        </xsl:if>
        <xsl:if test="not(@beginning_date='')">
            <coverage>
                <xsl:call-template name="temporalCoverage">
                    <xsl:with-param name="list" select="."/>
                </xsl:call-template>
            </coverage>
        </xsl:if>
        <attributeList>
            <xsl:call-template name="attributeType">
                <xsl:with-param name="list" select="entity_attribute_list/entity_attribute"/>
            </xsl:call-template>
        </attributeList>
    </dataTable>
</xsl:for-each>
```

direct mapping from nativeXML
test for empty entity description
call named template temporalCoverage
call named template attributeType

Figure 6. A partial example XSLT stylesheet to generate the EML dataTable element from native XML.

Summary

While there are many challenges in storing metadata in a RDBMS that begin with the determination of a RDBMS schema for storing metadata and include loading and updating metadata into the RDBMS, significant benefits can be gained. Using the RDBMS to store metadata in structured and relational tables enables flexible presentation of metadata, including the generation of EML. The described approach takes advantage of features within the RDBMS for generating native XML including the use of the data dictionary for creating a configuration file and programmatically interpreting the underlying database schema. The configuration file is used to help guide the process of populating native XML with attribute content from local RDBMS metadata tables. The XSLT stylesheets are used to complete the process of mapping the native XML into corresponding EML elements, but requires a fairly comprehensive understanding of both the local metadata tables and the EML schema. In addition to EML, other metadata document formats can also be easily generated using this described approach including PDF, HTML, Word, and other formats. Potentially, a customized script can be developed to map EML files back into the RDBMS, just the reverse process of the EML document generation described in this document, and we are exploring this possibility.

References

Center for Environmental Studies. 2002. "Xanthoria: A Distributed query system for XML encoded data", Arizona State University. Available on-line [http://ces.asu.edu/bdi/Subjects/Xanthoria/]

Costa, Duane. (2004). "EML Harvesting I: Metacat Harvester Overview and Management", LTER DataBits, Fall 2004 Issue, LTER Network Office. Available on-line [http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/04fall/]

Jones, Matthew B. (2003). "A brief overview of Ecological Metadata Language", LTER DataBits, Spring 2003 Issue, LTER Network Office. Available on-line [http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/03spring/]

Servilla, Mark, Brunt, James, San Gil, Inigo, Costa, Duane. (2006). "Pasta: A Network-level Architecture Design for Generating Synthetic Data Products in the LTER Network" LTER DataBits, Fall 2006 Issue, LTER Network Office. Available on-line [http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/06fall/]

Down the LTER Data Catalog Rabbit Hole

- Mark Servilla, James Brunt and Inigo San Gil (LNO)

Introduction

Have you ever wondered what you can find out about users of the LTER Data Catalog by looking at the inside of Metacat's database tables? Of course you have! Metacat's designers went the extra step to include an access log of who (actually, only the who's network address), what (the Metacat document's ID), and when (date and time) each metadata document is accessed. We began our investigation based on the need to better understand how metadata access is being tracked within the LTER Data Catalog in preparation for a similar task to monitor the access of LTER data as part of the LTER Data Access Policy and requirements. The following is a brief descriptive analysis of the Metacat "access_log" table. What we found is a story about one LTER site who's metadata is being accessed more than 70% more than all other documents in the LTER Data Catalog. Before plunging into this story though, we must raise the caveat that this is only a preliminary analysis of metadata access through the LTER Data Catalog, and its conclusions should not be used without further scrutiny.

Metacat's Access Log

Since its initial deployment at the LTER Network, Metacat has employed the use of an access log for collecting basic information when ever anyone accesses a metadata document in the LTER Data Catalog. Specifically, the log details the IP network address of the user, the access level of the user (registered user or public), the Metacat document ID, and the date and time of the event. In this analysis, we look only at "read" events. Unfortunately, our ability to perform a temporal trends analysis is limited to when the Metacat system was redeployed to new hardware at the LTER Network Office in July 2005. That's all right, though - the interesting information begins after July 2005.

First Light Down the Hole

The first query of the access log for "read" events netted a total number of 339,487 metadata document views for the period of July 2005 to March 2007. That's quite a lot of views for just under 2 years of operation (note that the LTER Metacat was actually deployed a number of years earlier on older hardware). Out of curiosity, we began to analyze how the network addresses were clustered around specific domains. To our amazement, it turns out that the majority of read events were from one specific domain - that of the Oak Ridge National Laboratory (ORNL). In September of 2005, and with the assistance of Dr. Inigo San Gil, the National Biological Information Infrastructure (NBII), as part of ORNL, began to harvest EML documents from the LTER Data Catalog to make them available from within their own data clearinghouse. The NBII harvest, however, did not account for all of the read events. Further investigation revealed that there were actually four primary domains that were reading documents from the LTER Data Catalog - (1) ORNL/NBII for harvesting to their data clearinghouse, (2) various search engine "bots" that apparently follow URL links that are set outside of the LTER Data Catalog, (3) internal users of the LTER Network Office for system testing and evaluation such as EML quality analysis, and (4) the general research community (Figure 1).

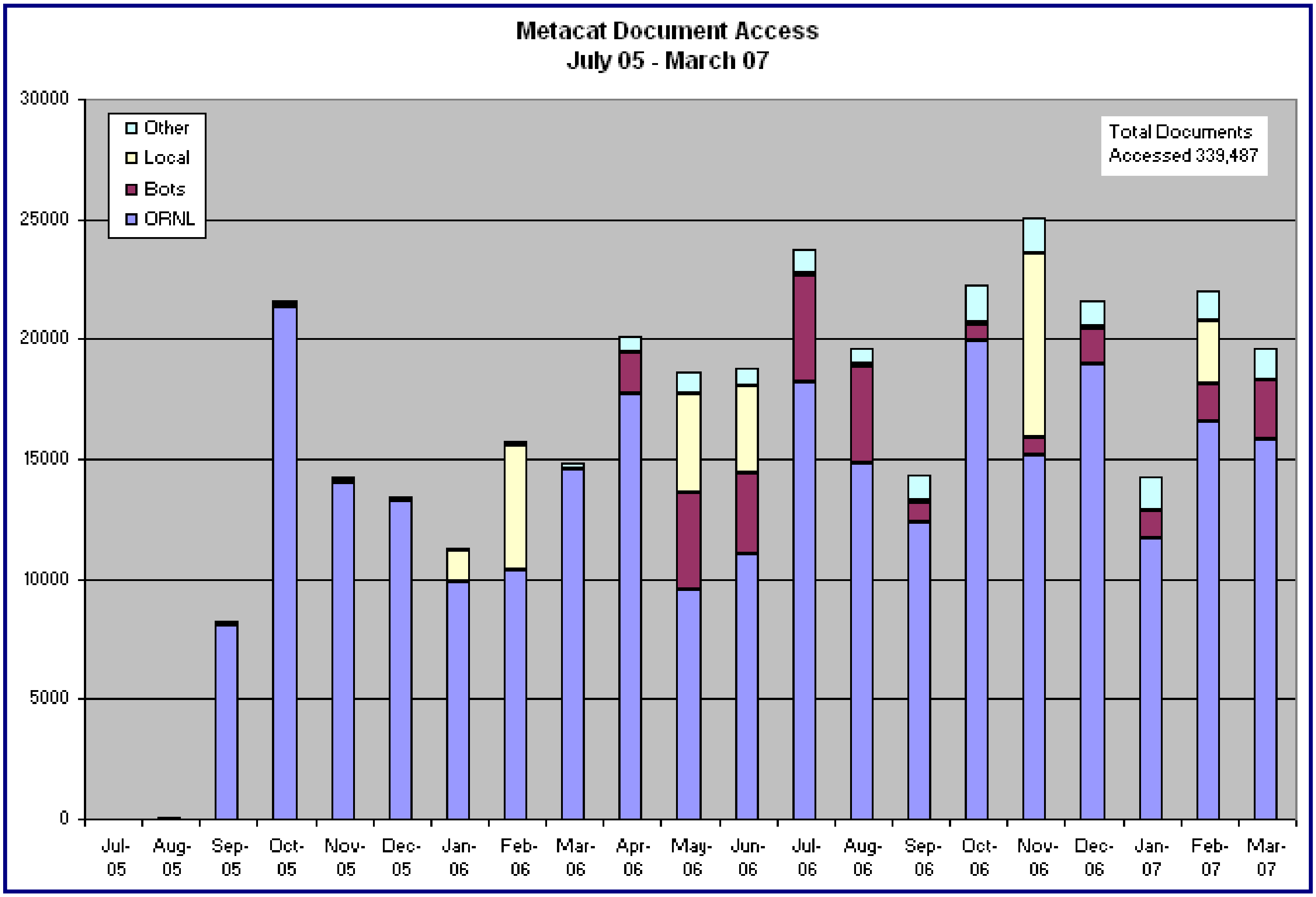


Figure 1. Histogram of all document "read" events from July 2005 to March 2007 broken into four primary categories - (1) ORNL/NBII, (2) search engine "bots", (3) local users from the LTER Network Office, and (4) other community users.

The Hole Gets Narrow

Although we awarded the "Jack La Lanne" prize to the NBII for exercising our Metacat system, (incidentally, the LTER metadata access stats at the NBII clearinghouse will be introduced in a future Databits article) we were really interested in how the general research community was accessing the LTER Data Catalog. To do this, we first filtered out all of the access log records from those domains that were of little interest. This left a set of 14,157 records (about 5% of all "read" events). Next, we partitioned the remaining records into one group of users who logged into the LTER Data Catalog through their registered accounts and another group who simply logged into the system as a "public" user. Again, we plotted this data as a histogram from July 2005 to March 2007 (Figure 2). Not surprisingly, we found that very few users actually logged into the LTER Data Catalog through their registered account. In other words, almost all users optioned to either not register or not log into the system even if they had previously registered.

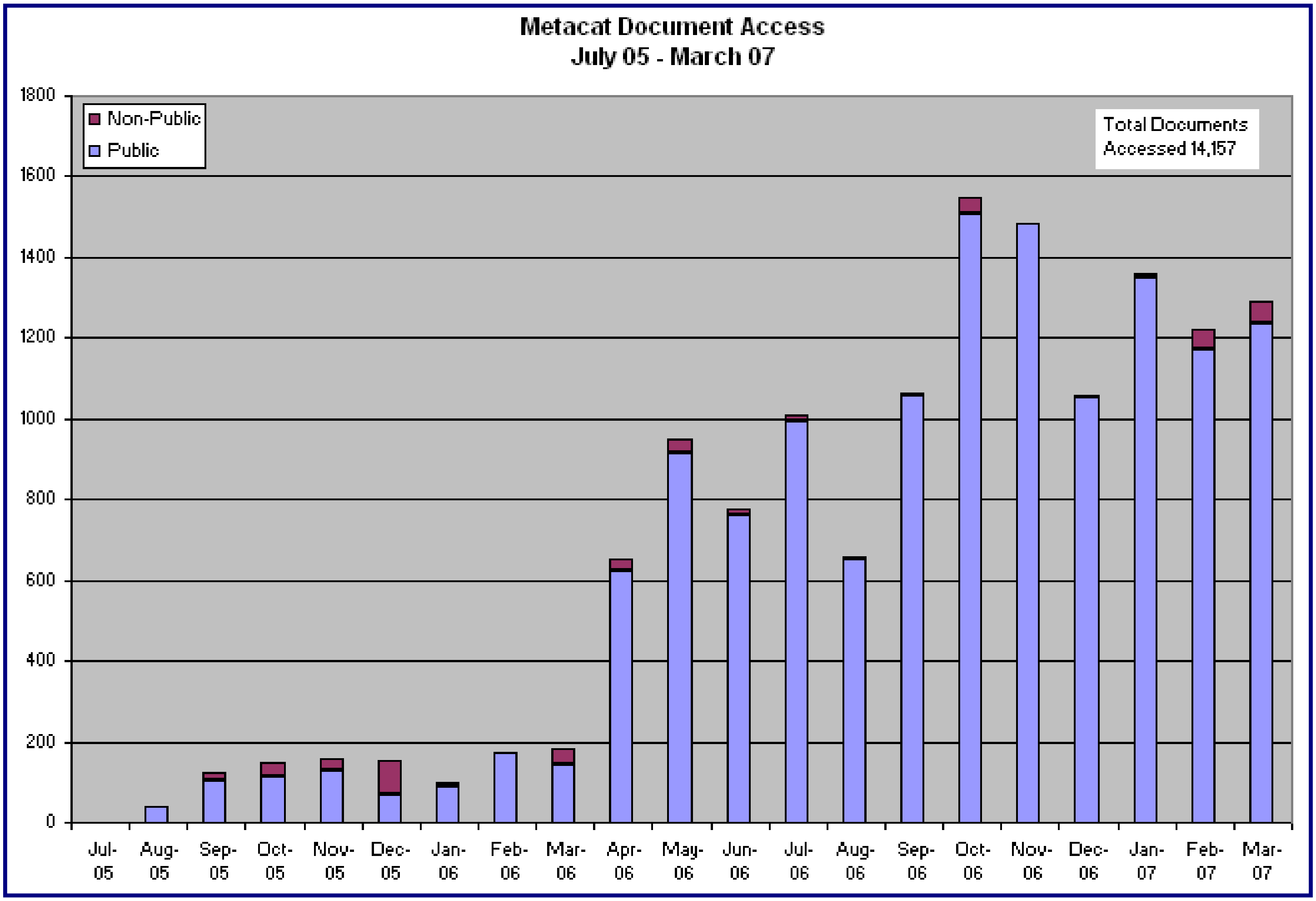


Figure 2. Filtered document access to the LTER Data Catalog showing only users of the general research community.

Alas, The Bottom of the Hole

We have finally reached the bottom of the rabbit hole. The good news illustrated by Figure 2 is that user access to the LTER Data Catalog is on the increase! But wait! There was a striking 3 times jump in access from March 2006 to April 2006. We naturally asked "what caused this marked change in use?" Now the task focused on picking apart the access log records for this period to see if any significant trends exist. What we found was quite startling (Figure 3). Prior to and including March 2006, very few documents that were accessed included those EML documents submitted by the McMurdo Dry Valleys LTER (MCM) site. This is plausible since MCM did not begin to harvest their EML documents until February 2006. What is most surprising is that after March 2006, greater than 70% (and generally between 80-90%) of all documents accessed through the LTER Data Catalog were from MCM. Well, it turns out that there are two additional factors besides the newly arrived metadata documents from MCM. First, MCM began to publish metadata from their website as fixed URL links to their EML documents in the LTER Data Catalog (this also coincides with the initial "bot" access events). The MCM website transitioned from offering their metadata details locally to placing links that served the metadata details using the LTER Data Catalog. Second, the LTER Network Information System development team also released the MCM version of the Metacat search skin (see 2.d in http://intranet.lternet.edu/archives/documents/reports/committee_reports/nisac/NISStatusReport20060515.pdf). This search skin provides identical functionality to the LTER Data Catalog query interface, but is displayed using the MCM style sheet and content. We recently provided a similar skin to the Baltimore Ecosystem Study LTER site, and skins for the Sevilleta and Luquillo LTER sites are in the works.

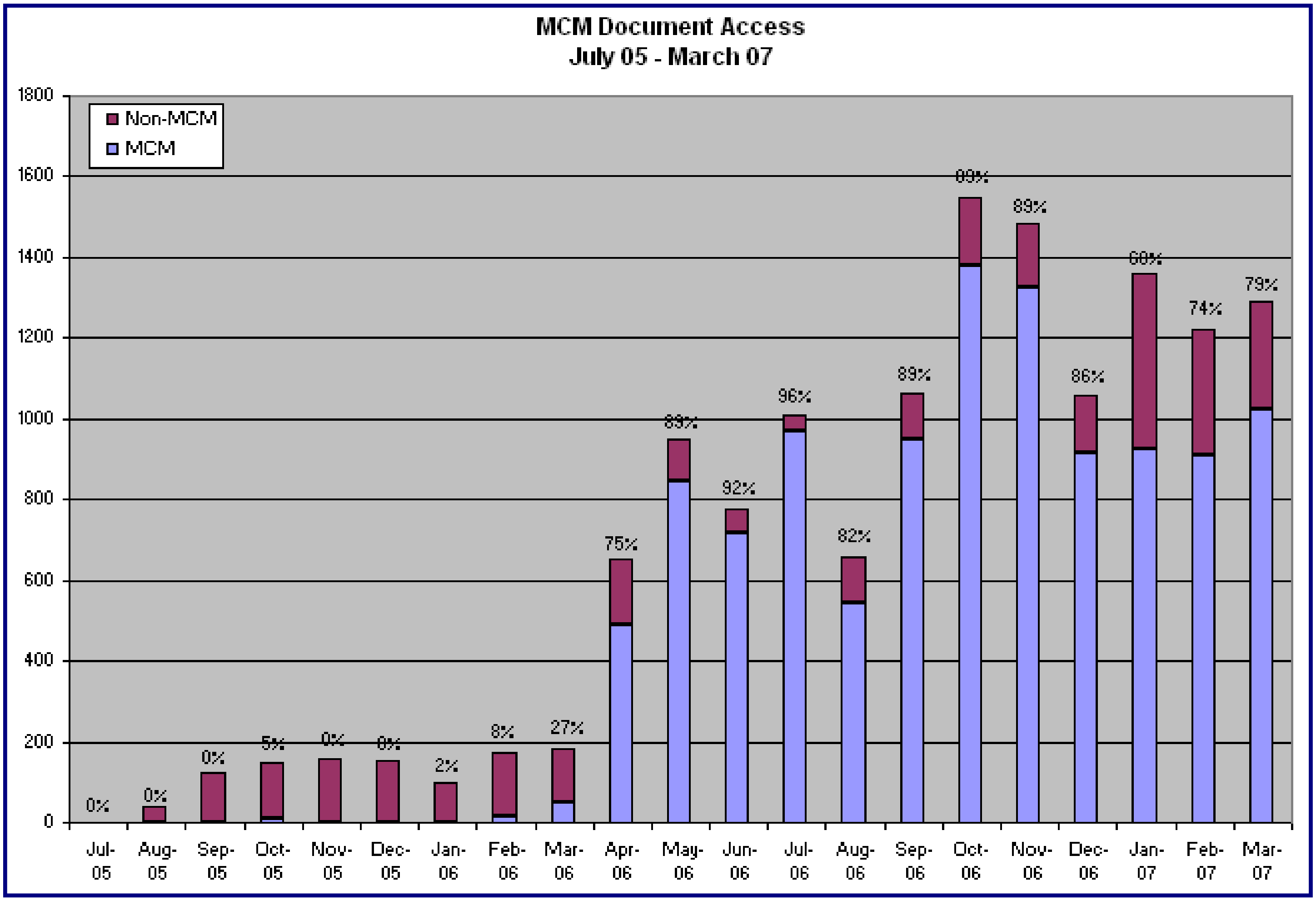


Figure 3. Distribution of MCM document access compared to overall document access for the general research community. Percentage values above each histogram represents the MCM documents that were accessed.

There could, however, still be yet another rock in the hole that was throwing off our interpretation of the data. "What was the distribution of the MCM document access?" "Was it from a single domain or sub-domain of users?" That is, what if there were other automated systems on the Internet that were querying only MCM documents? It turns out that a small percentage of the documents accessed were done so by the Ohio State University (OSU) domain, current home of the MCM information management team. For example, 146 documents were accessed by the OSU domain in April 2006 and 65 documents in May 2006. Most users (or should we say network addresses) only accessed between 1 and 2 documents during the month period. As we can see in Figure 4, the number of unique network addresses accessing a larger number of EML documents diminishes quickly. From this, we infer that the user base is fairly well distributed across the community. In other words, there appear to be lots of users accessing a small number of documents as opposed to a few users accessing a lot of documents.

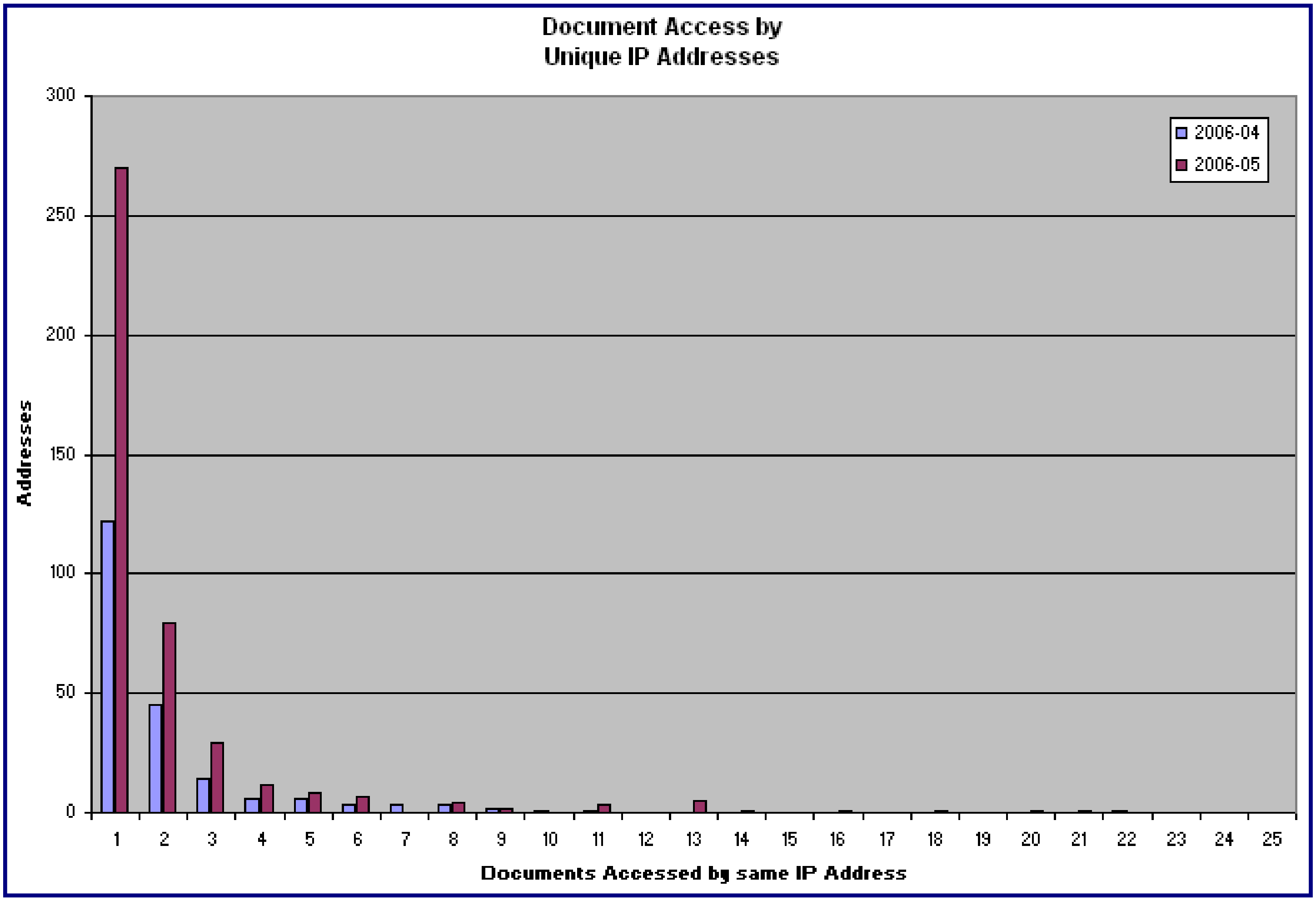


Figure 4. Number of unique IP network addresses that access one or more metadata documents in the LTER Data Catalog.

Conclusion

Looking up from the bottom of the hole, it is clear that the LTER Data Catalog (and the underlying Metacat system) are being used on a regular basis by various end users - some automated, some not. After peeling away the automated and test users from the data, what becomes evident is that documents from the MCM LTER site have the highest access rate. Although we have speculated as to the reasons behind this, it is likely because MCM relies heavily on the LTER Data Catalog to provide metadata management for their project. All links for their metadata point back to the LTER Data Catalog. By leveraging the LTER Data Catalog, MCM can redirect resources to other aspects of their project.

Getting back to our original goal of understanding what the access log data has to offer, we find that we can easily provide MCM (and the LTER Network sites) with metadata access reports based on information held within the LTER Data Catalog. Building on this approach, we envision a similar system that will provide for the provisions in the LTER Data Access Policy. We are now in the process of preparing an RFC (Request for Comments) on our proposed approach to gain feedback from the LTER information management community on its efficacy.

Editorials

A Web Developer’s View of the Research World and the Entertainment Industry

- Shaun Haber (PAL & CCE)

Having worked with two LTER sites for a number of years, I left to pursue web development as a profession within the entertainment industry. This provided me an opportunity to compare the work of a programmer within the academic world with that in the entertainment realm. The two industries are similar in many ways. Granted, there are some obvious material differences. The entertainment industry is high-profile and generally pays better than academia, but academia tends to be a "quieter" working environment.

The similarities? A web developer’s work in academia typically involves the same core set of duties and responsibilities as in entertainment. More interestingly, a web developer in both industries experiences the same social disconnect among peers and co-workers unfamiliar with the technical "know-and-how". However, academia may offer a calmer environment for bridging the communication gap caused by the technical polarity.

I can safely say this because I’ve worked in both arenas. My background begins in 2004 at the Scripps Institution of Oceanography, where I joined Karen Baker and Jerry Wanetick in building a social framework to encourage the participation, collaboration, and the exchange of scientific data and ideas among various faculty, staff, and students. Our local vision took a grounds-up approach as we examined ways to improve our technical infrastructure. (This was an opportunity for me to experiment with various open-source content management systems, eventually settling with the blogging platform WordPress.)

We grew a small community with engaged participants. We called our community "Ocean Informatics", however, it evolved into something more. Ocean Informatics became an abstraction where a programmer (i.e. myself) and a scientist (i.e. someone else) could productively discuss ideas involving the integration of scientific data and technology. One example involved creating the SCCOOS data schema, where experts from varying fields (oceanography, information management, computer science, etc.) worked together in building a relational database that would store millions of records of streaming data from moorings.

In 2006, I had the fortunate opportunity to take a technology project management role at a major recording label in Burbank. Saying goodbye to Scripps was hard, but I was eager to move forward with my career and face new challenges in a new environment. In some ways, my job hasn’t changed. Working in the New Media department, which spans technology and marketing, I use the same technologies (PHP, MySQL, Subversion, etc.) as before. I manage the same social disparity between the technical and non-technical people.

A scientific researcher in academia is very like a marketing director in entertainment with respect to the internet. Both depend upon a variety of technologies to expose their work to a larger respective audience. (A scientist publishes data to the scientific community, and a marketing director promotes a band to its core demographic). Essentially, they both embrace the openness of the internet (HTML, XML, APIs, etc.) to publish data and share content. This requires a strong knowledge of the latest technological tools and trends.

Though we exchange plenty of ideas in the New Media department, involving marketing and the internet, the industry often moves extremely fast and offers limited time for grass-roots participation. It is a business, after all, and has a very different primary need than an academic institution, which places more value on research and experimentation. The music industry has no time or interest in supporting a meta-community that encourages collaboration from people of various fields and that studies itself. If it did, maybe we could call it "Music Informatics"?

With Ocean Informatics, I can stay connected with the community, despite lacking a physical presence. This is beneficial for everyone involved. I can occasionally contribute my time, knowing my input will be well received. This includes posting to a blog, participating in an online video chat, or coming down in person to help troubleshoot technical issues (e.g. last month I helped a PI investigate a software bug).

It is optimistic to assume the technical polarity I have experienced is weakening as more people grow comfortable with computers and learn to embrace the latest technologies to fit their needs. However, there still exists a great disparity between technical and non-technical people in any industry. It is important to realize this is a persistent, global concern, and not only limited to science. Finally, it is noteworthy that an academic institution provides an environment suitable for studying this communication gap and simultaneously shrinking it.

Preparing for your LTER midterm review

- Jonathan Walsh (BES)

I hope this article will serve as a helpful guide for LTER information managers to prepare for a site review. It is important to recognize that each review team includes at least one member who is an information management expert.

Help plan the presentations

Your Investigators are undoubtedly ready to present their work. It will be helpful if you convince them to show you their slides and their outlines so that you can see how the flow of information from their work fits into your scheme for managing data. See that you are cognizant of their system of data collection and storage. Be sure that you are aware of the data they archive and that, if the site and LTER data policies permit, these data are being made available to the public.

Ensure the technical details are in order

As the information manager you might be helpful in ensuring the presentations go smoothly on the big day. Things like projectors, screens, extension cords and flash drives are necessary and you might want to take on the role of making sure all these things are in place. An LCD projector bulb, for example, is fiercely expensive and virtually impossible to find on short notice.

A list of details you might help with includes:

- Seating
- Wireless connectivity, to permit the review team to scan your web materials during the meeting
- Projector/backup projector/Screen
- Extension cords
- Grounded outlet adapters
- Flash drives
- Spare laptop for presentations

Label the file names for the talks so they sort in order of presentation, or are otherwise easy to find in the seconds before a speaker starts talking.

Practice - rehearse

Go to the place where you will be presenting and set everything up. Make sure it all works.

Hopefully, your entire team will do a rehearsal. If so, use it as a chance to test all the technical details. Assuming your site plans to have a series of talks, this rehearsal will let you see how the presentations transition from one to another, with speakers moving up to the podium and back to their seats, loading the media onto the projector, and so forth.

Post a wiki for the review team

Set up a wiki for the review team and post information they might need such as the agenda, logistics, maps and the like. A wiki is also a great place for the team to find the handful of representative or key reprints they will ask for, and perhaps for posting your site annual report or other information they may request.

Consider Giving an information management presentation

Ask your lead PI or the committee that sets the agenda if you can give a presentation about your own work as the information manager. Gear it mostly toward a non-technical group. In a 15 minute talk you should be able to communicate the following points. Note the emphasis on showing the team things, not just mentioning them.

Show how you plan collection and management of information for new studies.

Show how your data are kept. Describe how you archive and protect data. Describe the machines and systems that are involved. Be sure to show how your data are available to the public as well as to scientists within your site.

Show how you generate metadata. Describe the life cycle of the metadata beginning with its structure – what fields are collected, how it is collected, where it is collected, and what forms it is then presented in to make it useful to the scientific community, LTER and beyond.

Show how you assess and assure quality. Describe as well your means for keeping data backed up.

Show how you collaborate. Make it clear that your data are useful and available as part of larger group activities. In the case of LTER data that means that they can be found on the Metacat server by means of your metadata stream expressed in Extensible Markup Language (XML). But also show how your data are available for collaboration and synthesis with entities beyond the LTER network. Make sure you describe the nature of the LTER network. It is intended to be a loose collaboration of scientists, and not a sensor network. Your audience may not be aware of the intricacies in information management that arise as a result of such a design and the unique ways the LTER has developed to collaborate.

Show how you participate in LTER network activities. Describe the committees you serve on. And talk about how you participate in the LTER information management committee annual meeting. Again, several of the persons who review your site are likely not involved with the LTER and they need to understand your role and your mission as well as how it fits in with those of your LTER colleagues.

Get your website updated

Your website is your portal to the public. Although obvious, it's mentioned here because your review team will undoubtedly go over it carefully before the review. Make sure your research products and data are easy to find. Try to have it completely up to date before the review team is chosen because some reviewers might visit your website the minute they learn they will be reviewing you. If you update it at the last minute, they may never see important new material.

Have an "elevator" speech

You might just have a few precious moments with your information management reviewer so you should have a concise, friendly, professional dialogue in mind. You must also be keen to find out what concerns the reviewer has and be prepared to address them.

News Bits

Information Management Committee: GIS, Technology, and Changing Organizational Structures

- Nicole Kaplan (SGS), Corinna Gries (CAP), Karen Baker (CCE & PAL), Don Henshaw (AND), Theresa Valentine (AND) and John Vande Castle (LNO)

Recently, the LTER Network Office (LNO), the Information Management Committee, GIS working group, LTER Technology Committee, and LTER Executive Board (EB) explored the potential for broadening the focus of the IMC to include the activities of two other groups, the GIS/RS working group and the Technology Committee. This was driven by several factors including the considerable overlap of focus and participants among committees, the availability of resources for face to face meetings, new video conferencing tools, recent dips in activity in several committees, and the perception that the different committees were working independently rather than interactively. The discussion considered that this broader committee focus would better allow for achievement of integrative science as put forth in the ISSE (Integrative Science for Society and Environment, A Strategic Research Plan), and allow for coordination of activities across these groups. The suggestions generated a lively discussion across the community of information managers, GIS and IT professionals considering governance structures, group identities and improvement of collaboration among the diverse specialists.

Although the scope of this article is not the history of LTER information management, we will give some background information on the different groups of interest before discussing the new challenges and opportunities we are facing. LTER Data Management has a long and interesting history which can be explored in reports from workshops and meetings dating back to 1988 (see Shugart 1988, Gosz 1989, Foster and Boose 1991). Reviewing these documents provides insight into how information management and the implementation of new technologies have developed to meet the needs of diverse ecological research projects. The digital arena was informed by many early works including research in Scandinavia on information systems (Langefors, 1973) and in the US on computer systems (Friedman, 1989; see Good Read this issue). By the late twentieth century, professional opportunities were being identified for work with data and information. Research Information Management (RIM) was becoming recognized as a scientific discipline with curricula at the BS, MS and PH.D level combining biology and ecology with computer science. (Andrews et al. 1992; Abbott 1988). Reviewing the old reports reveals how within the LTER system, working groups organized around specific information technology related subjects. Early on these committees were formed ad hoc and with changing membership according to the subject under consideration until LTER implemented greater formalization for some committees (LTER 1994).

After a phase of organization that started in the early 1980’s and continued with the establishment of regular annual meetings, the information managers developed a vision statement (LTER 1995, 2001) and were recognized as a standing network-wide Information Management Committee (IMC) by the Scientific Steering Committee of the LTER network. All LTER Standing Committees provide regular reports and recommendations to the LTER Network Science Board. By the late 1980’s a minimum standard installation for sites was under discussion which developed into the establishment of a working group to design and develop the Network Information System (NIS; Baker et al 2000). This working group developed a framework for cross-site modules including the personnel directory, publication bibliography, site description directory, and site climate database. In addition to addressing committee organization, data sharing, metadata, and communication issues, the committee worked to establish the Network Information Advisory Committee as a forum for dialogue about research needs and information management.

In conjunction with the growing role for technology within a distributed network such as the LTER, a Technology Committee self organized and was established as a standing committee in the 1990's to provide leadership on applications of technology within the LTER Network and to seek out new technologies that would advance ecological science. The Committee considers existing and emergent technologies and attempts to define new measurement needs where no applicable technology exists. A list of recommendations was developed, many of which have been implemented over the last few years, including direct internet connectivity at LTER research sites, training in, adoption and deployment of wireless communication and data transmission with various sensor deployment by a number of LTER sites, and important computer infrastructure for support of LTER Information System implementation. The Technology Committee initiated registration of all sites for high resolution reconnaissance data with the Global Fiducial Program and gathered new information related to MODIS time series data, more regular photographic surveys from the International Space Station, and LIDAR data for LTER sites.

Another group, focusing on Geographical Information Systems (GIS) was developing throughout the 1990’s. GIS technologies, including Global Positioning Systems (GPS) became available to some sites. At annual workshops and meetings GIS software and hardware were demonstrated and GIS data requests continued to increase at a steady pace. A Network GIS/RS lab was made available for ecologists to perform analysis on different LTER sites using satellite imagery and training workshops were organized. GIS coverage and access to GIS data grew (Adams et al. 1995). The IMC recognized the importance of spatial data with respect to information management activities and hosted a one-day geospatial data symposium in Seattle in 1994 inviting speakers to the annual meeting. GIS professionals across the network met several times and reported GIS issues related to LTER Network Information System development to the IMC. It was recognized early on that in the area of data management the IMC and GIS working group had similar needs and problems. However, the groups were not working together enough for sharing ideas, solutions, and development costs (Valentine 2002).

This brief excurse into the development of Information Management and Technology at LTER shows the degree of overlap in focus of the three committees. However, it also makes clear that increasingly expert knowledge in the areas becomes essential for successfully meeting future challenges. Accordingly, in meeting the challenge to better integrate efforts we are considering the proposal to add members from the GIS/RS and technology leadership to the Information Manager's steering committee (IMExec) and expand the next annual IMC meeting to include representatives from the other two committees. One focus of the meeting will be GIS/RS data management. At the meeting discussions will take place regarding structures for efficient communication among the now larger number of participants, meeting venues to associate with groups other than the ESA, increased exposure of LTER IM to the different science disciplines and collaboration with LTER scientists.

The 20-year NSF review of LTER (NSF 2002) brought greater attention to the need for synthetic activities; the LTER community designated the third LTER decade as the "Decade of Synthesis". The information management community is now challenged with supporting cross-site and synthetic research and integrating diverse datasets over various platforms. Accordingly, there is a strong interest to coordinate existing and advanced technologies across LTER sites and a need to develop common approaches to integrating and analyzing new and traditional types of information holdings. The management of a variety of data types – from field study data to GIS and remote sensing data as well as, but not limited to, streaming sensor, high volume analytical instrument outputs, and social science data - will continue to play an important role in cross-site, regional, and global data synthesis and analysis.

In the coming years, the members of the broader Information Management community will be challenged to design practices and cyberinfrastructure to incorporate and integrate new data and technologies. Our community has already been successful in collaborating as members of interdisciplinary teams and bridging across interdependent science, data and technology issues. The expansion of IMC responsibilities to include spatial data issues and sensor technologies will demand better planning and communication with other informatics and science disciplines. New joint efforts will create opportunities to foster communication across fields related to data as well as to actively develop new collaborative approaches to information management and infrastructure building (Baker and Millerand, 2007) as concepts, strategies, technologies, and research questions are identified, formulated, and then change.

References:

Abbott, A. 1988. The System of Professions: An Essay on the division of expert labor. The Univ of Chicago Press, Chicago.

Andrews, M., K. Baker, B. Benson, E. Boose, J. Brunt, C. Bledsoe, J. Briggs, G. Calabria, A. El Haddi, D. Henshaw, R. Ingersoll, T. Kirchner, M. Klingensmith, L. Krievs, J. Laundre, R. Lent, E. Melendez, W. Michener, B. Nolen, R. Nottrott, J. Porter, S. Stafford, C. Veen. 1992. Proceedings of the 1992 LTER Data Management Workshop. (http://intranet.lternet.edu/modules.php?name=UpDownload&req=viewsdownload&sid=6&min=10&orderby=dateD&show=10)

Adams, P., K. Baker, B. Benson, D. Blodgett, J. Briggs, C. Bristow, J. Brunt, G. Calabria, H. Chinn, D. Chang, M. Easter, J. Gosz, M. Hartman, J.Hastings, B. Hayden, D. Henshaw, R. Ingersoll, T. Kirchner, K. LaFleur, R. Lent, E. Melendez, B. Nolen, R. Nottrott, J. Porter, S. Stafford, M. Tomecek, J. Vande Castle, C. Veen. 1995. Proceedings of the 1995 LTER Data Management Workshop. (http://intranet.lternet.edu/modules.php?name=UpDownload&req=viewsdownload&sid=6&min=10&orderby=dateD&show=10)

Baker, K.S., B. Benson, D.L. Henshaw, D. Blodgett, J. Porter and S.G. Stafford, Evolution of a Multi-Site Network Information System: the LTER Information Management Paradigm, BioScience 50(11):963-983, 2000.

Baker, K.S. and F. Millerand. 2007. Articulation Work Supporting Information Infrastructure Design: Coordination, Categorization, and Assessment in Practice. Proceedings of the 40 th Hawaii International Conference on System Science. Pp 1530 – 1605. http://cce.lternet.edu/docs/bibliography/026ccelter.pdf

Foster and Boose 1991. Technology Development in the LTER Network: http://intranet.lternet.edu/archives/documents/reports/Technology-reports/techdev1991.html

Geospatial data symposium in Seattle 1994 report: http://intranet.lternet.edu/archives/documents/reports/committee_reports/Data-management-committee/1994-DM-committee-report/im_1994_report.pdf

Gosz et al. 1989. LTER Technology Report for the LTER Strategic Plan: http://intranet.lternet.edu/archives/documents/reports/Technology-reports/goszrpt.html

LTER 1994 Fall Coordinating Committee Meeting Minutes: http://intranet.lternet.edu/modules.php?name=UpDownload&req=viewsdownload&sid=1

LTER 1995 Data Management Workshop: http://intranet.lternet.edu/archives/documents/reports/committee_reports/Data-management-committee/1995-DM-committee-report/im_1995_report.htm

LTER 2001 Ecoinformatics Vision Statement: http://intranet.lternet.edu/archives/documents/vision_statements/ECOINFORMATICS.htm

LTER Intranet 2006. LTER Intranet webpage: http://intranet.lternet.edu/

Michener and Stafford, 1988. Report to LTER Coordinating Committee: 'Summary of LTER Data Managers' Workshop: http://intranet.lternet.edu/archives/documents/reports/committee_reports/Data-management-committee/1988-DM-committee-report/1988DMreport.pdf

NSF 2002 20-year review of LTER report: http://intranet.lternet.edu/archives/documents/reports/20_yr_review/

Shugart et al. 1988. Report Of The NSF Advisory Committee On Scientific And Technological Planning For Long-Term Ecological Research Projects: http://intranet.lternet.edu/archives/documents/reports/Technology-reports/Shugart/shugart.htm

Valentine, T. 2002. LTER GIS Meeting Summary. http://www.lternet.edu/technology/ltergis/working_group1/index.html

Strategic Mentoring for LTER IMs

- Susan Stafford, University of Minnesota

Over 30 participants, decided to forgo free time exploring the Colorado Rocky Mountains at the 2006 All Scientists Meeting (ASM) to listen to Susan Stafford, University of Minnesota, and Peter Arzberger, UCSD University of California, San Diego - who joined the group via skype – to discuss opportunities for long-term mentoring and professional development for LTER Information Managers (IMs).

NSF's Cyberinfrastructure (CI) vision (developed by the Office of Cyberinfrastructure) will revolutionize science and engineering through cyberinfrastructure. The CI vision framework contains several components: high performance computing, data, data analysis, and visualization, access and control of remote instruments including sensors, virtual organizations, and learning and workforce development. Implementing the CI vision will include establishing a portfolio of applications, forming collaborative teams, cultivating innovative uses, and fostering the growth of a community of biologists that use High Performance Computing resources.

The premise of this workshop was the arrival of CI creates the "perfect storm" for strategic mentoring opportunities for LTER IMs. The intersection of the new research paradigm for the LTER Network (as outlined in the LTER Strategic Plan) with the new research initiative Integrative Science for Society and the Environment (ISSE) and the advent of (CI) creates unprecedented opportunities for the IM community to promote integrative, transdisciplinary network science.

As the LTER embarks on its third decade, CI will be embedded within the LTER network in unprecedented ways. CI is fundamentally changing how LTER Science will be done. CI will be "where the action is" for the next decade. This includes sensors, sensor networks, and observing networks in general. CI and *e-science* are bringing resources to researchers which include computing resources for modeling, data analysis, and interfaces for visualization and collaborations, digital collections for knowledge management, and instrumentation for observations, all enabled by CI to create a globally connected human community.

What distinguishes this new era of *e-science* is the interconnectedness of four forces: persistent infrastructure, enabling technologies, educational and capacity building, and sustained collaborations - all working together to achieve previously unobtainable observations and analyses. *E-science* recognizes all four forces as necessary - and in fact essential – if *e-science* is to achieve it’s full potential.

Several examples of successful global team science were discussed:
PRAGMA: http://www.pragma-grid.net
PRIME: http://prime.ucsd.edu
PRIUS: http://prius.ist.osaka-u.ac.jp/en/index.html
GLEON: http://www.gleon.org
OptIPuter: http://www.optiputer.net

To be successful, each had to address technical, sociological, teaching and education, standardization, financial, and political issues.

Opportunities for strategic mentoring and professional development will come from these expanded partnerships. Several were discussed which included:

- Pursue rotations at NSF within the directorates of BIO, INT, SBE, and OCI
- Work with the computer science community to write joint proposals
- Find opportunities for internships and writing joint proposals with industry partners
- Reach out and include social scientist colleagues in joint proposals
- Don’t ignore the policy makers – invite them to join collaborations
- Get connected internationally

Bottom line - reach out and make meaningful linkages to insure the LTER IM community forges successful international, e-science collaborations. IMs must seize these unprecedented opportunities and use them as the vehicle for advanced mentoring and professional development opportunities. In summary, this "perfect storm" will either result in an oncoming tsunami or the surfer’s wave of a life-time. The choice is up to the individual IM and their site leadership. But don’t take too long to decide – the window of opportunity is closing quickly so act accordingly.

LTER Site GIS Survey Results

- Barbara Nolen (JRN)

In response to the e-mails that have been circulating regarding the GIS committee, I was interested in looking at the results from the 2003 GIS survey and finding out who is involved in GIS at each site currently. I was able to access the surveys from the13 of 24 LTER sites that responded. But as the survey is already 4 years old I wanted to find out the current responses to the survey. So I summarized the results of the 2003 survey by quickly going through the surveys and simplifying the results.

From this simplified view of the 2003 survey, I produced a quick telephone survey for 2007 and called each site and spoke to the GIS person and tallied the results.

Here is my approach to the simplified GIS survey for 2007:

GIS person - **Y** if the site has a dedicated GIS person
N if the site has an IM that also handles GIS

ESRI -	**Y** if your site uses ESRI products for GIS **N** if your site does not use ESRI products for GIS
Boundary -	**Y** if the site has a shapefile, coverage, or simply coordinates that define an area **N** if the site has no GIS representation
Study Locations -	**Y** if some study sites locations are included in GIS **N** if no study site locations are included in GIS
DEM -	**Y** if the site has access to DEMs or other terrain products **N** if the site has no access to terrain data
Landsat -	**Y** if the site has access to Landsat or other satellite products **N** if the site has no access to satellite data
DOQQ's -	**Y** if the site has access to DOQQ's or other aerial products **N** if the site has no access to aerial data
Core sites -	**Y** if some core sites locations are included in GIS **N** if no cores site locations are included in GIS
Online -	**Y** if the site has some spatial data available online **N** if the site has no spatial data available online
EML -	**Y** if the site is using EML **N** if the site is not using EML at all
GPS -	**Y** if the site has access to gps equipment **N** if the site has no access to gps equipment
Internet Map Site -	**Y** if the site has a map online the can be used for information **N** if the site has no map online
Integrating tab/spatial -	**Y** if the site is working to integrate tabular and spatial data **N** if the site is not currently working to integrate tabular and spatial data

I did not address remote sensing or spatial analysis/modeling in this survey. I also asked each site if the GIS (ESRI) software was cost prohibitive or readily available through a university site license.

Amazingly I was able to contact 26 of the 26 sites for current answers to this slimmed down survey.

The original 2003 survey that this survey was based on was developed with careful consideration and input from most of us involved in the GIS workshop in 2003. We considered it to represent 'minimum capabilities" for a site. Even though not all sites responded in 2003, many, if not most, were aware of the survey efforts and questions.

2007 GIS telephone survey summary:

All sites:
- use ESRI software for GIS (even oceanographic sites use it to supplement Matlab or IDL)
- have boundary layer, study locations and core sites
- have access to gps equipment (although Gil's not sure)
- have access to ESRI software through a university site license (although Eda's not sure)

12 sites: have a dedicated gis person

23 sites: use DEMs and Landsat

18 sites: use DOQQs (keep in mind they are not available for all sites)

22 sites: have some spatial data online

25 sites: are using EML (In IM not necessarily meshed with GIS)

18 sites: have an internet map site

20 sites: are working to integrate tabular and spatial data at some level

I feel great progress has been made at each site to meet these minimum capabilities.

Survey participants and additional interested people:

AND – Theresa Valentine
ARC – Andrew Balser
BES – Jonathan Walsh
BNZ – Brian Riordan
CAP – Corrina Gries
CCE – Karen Baker
CDR – Dan Bahauddin
Troy Mielke
CWT – Barrie Collins
FCE – Mike Rugge

GCE – Kris Meehan
HFR – Brian Hall
HBR – John Campbell
JRN – Barbara Nolen
Ken Ramsey
KBS – Suzanne Sippel
KNZ – Jincheng Gao
LUQ – Eda Melendez
MCM – Christopher Gardner
MCR – Will McClintock
Margaret O’Brien

NTL – Jonathon Chipman
Barbara Benson
Jeff Maxted
NWT – Todd Ackerman
PAL – Karen Baker
PIE – Gil Pontius
SBC – Margaret O’Brien
SEV – Kristin Vanderbilt
SGS – Nicole Kaplan
Bob Flynn
VCR – John Porter

We need to keep in mind that across sites there are quantitative differences between sites with regard to the number of researchers/research sites that are supported and the number of spatial data tasks that are the responsibility of the GIS person(s).

I hope this exercise offers information as to who the primary GIS contacts are at this point in time for each site and where the collective sites are with regard to the survey questions that we used in 2003.

I have just completed another remote sensing telephone survey that will be summarized in the next Databits.

On-going research collaboration

- Florence Millerand (PAL & CCE)

I have been working as a postdoc researcher associated with LTER within a social science research project with Karen Baker, Geof Bowker and David Ribes, the Comparative Interoperability Project (http://interoperability.ucsd.edu/; see Strategies for Building Scientific Cyberinfrastructure, LTER Network News Spring 2005, p17, http://www.lternet.edu/news/images/spring05/NetworkNewsSpring05.pdf). I am now working at University of Quebec, Montreal (UQAM) as a faculty in the Communication Department (http://www.dcsp.uqam.ca/) with continuing ties to UCSD and LTER. Research and teaching responsibilities began in Spring of 2006. In brief, Communication Studies is a research field in the social sciences that brings an interdisciplinary approach to study the interactions between people, groups, and organizations as well as relations to medias and information technologies productions and uses. University breaks have permitted two return visits to

San Diego in 2006- 2007 for continued field work and research collaboration and will permit attendance at the upcoming summer LTER Information Management Meeting.

At UQAM I am actively involved in an interuniversity research center on science and technology (CIRST: Centre interuniversitaire de recherche sur la science et la technologie, http://www.cirst.uqam.ca/English/tabid/94/Default.aspx). CIRST is Canada’s primary research group dedicated to the social study of scientific and technological activity, and is characterized by a broad range and critical number of associated researchers, a high level of interdisciplinarity and its interuniversity arrangements. The center groups together more than 40 regular and associate members from nearly ten institutions of higher education, most of them located in the province of Quebec.

The CIRST research is organized around three main lines:

1. Scientific and technological development analysis
2. Socio-economical analysis and management of technology
3. Socio-political analysis of technology uses and effects

Research at CIRST is aimed at furthering knowledge and using it to make a contribution to the elaboration and implementation of policies, as well as resolving social problems that carry scientific or technological dimensions.

Research collaboration is ongoing with PAL and CCE sites, pursuing ethnographic work on the enactment of standards, LTER information infrastructure initiatives as well as on Ocean Informatics developments at Scripps Institution of Oceanography – all efforts to improve knowledge and understandings of interdependent organizational and technical changes in order that practice may better inform developing research arrangements. Future research projects include a research monograph of Ocean Informatics from an action research perspective as well as a comparative study between Canadian and US LTER collaborative research networks.

Good Reads

Computer Systems Development: History, Organization and Implementation

- Karen Baker (PAL & CCE)

Andrew L. Friedman, Computer Systems Development: History, Organization and Implementation. John Wiley, New York, 1989.

As information managers we use computer systems in our day-to-day work. This book provides background for our dynamic digital data environment and tools. Unlike many classic computer systems texts, the author highlights non-linear aspects of computer systems development, taking a historical rather than prescriptive perspective. The historical perspective provides insights that inform today’s work. The history starts with business applications created in the 1950’s, develops under the influence of legacy systems, and evolves together with technological and social factors. Friedman states: "If we are interested in understanding how things are likely to change, it is important to understand how things came to be as they are." In the information world, this background material is pertinent to the work of many: computer programmers and information managers, computer center directors and network engineers, technologists and data analysts, researchers and policy makers.

The book is organized into four sections: 1) Background and theory; 2) Phase one and phase two: hardware and software constraints dominate; 3) Phase three: user needs dominate; 4) Prospects and implications. The concept of phases introduces the notion of change and planning over time. In the first part, a rationale for a historical and non-prescriptive approach to computer systems development is given along with a guide to readers. This is followed by an overview of previous approaches to the history of computer systems and the book’s own phase model of computerization growth. The second section after discussing hardware and software constraints, introduces strategies for dealing with the software bottleneck. The third section summarizes successful systems development, life cycles, and a phase where user needs dominate. It asks the non-trivial question: Who are the users? The literature on user relations is reviewed and strategies for dealing with user relations' are discussed. Throughout the book the concept of 'Agents of Change' is explored. It is noted that each of the approaches to the history of computer systems development has a different fundamental change agent, "a different factor which may be thought of as driving the history, of stimulating long-run changes".

The book concludes with discussion and evaluation of some models of computer systems development. In addition, a set of generalizations from the Phase Model perspective are given: 1) no 'best way'; 2) auto generation of new technology; 3) perception of problems; 4) technology regimes and computer systems technology regime; and finally 5) directions of causality. In the final generalization, both technological and social factors are considered.

For information system designers, technology enablers and end-users, this book provides context – both historical and synthetic. Today we face the difficulty of hearing about techniques implemented that are not widespread but rather exist at a few state-of-the-art sites. Techniques may be understood as widespread, when the techniques may be described more accurately as potentially widespread or as hoped-for ubiquitous applications. Indeed, it is interesting to note over a two-decade uptake timeframe, the concept of "Agents of Change" traveled from the community - as mentioned by Friedman in this 1983 book - to become in 2004-2007 an NSF Human Social Dynamics funding call proposal category of 'Agents of Change'. Early histories point to hardware, applications, and the labor process as fundamental change agents. Friedman underscores a variety of types of change factors as well as the need to modify recognized drivers.

Friedman outlines in the foreword what he sees as dimensions of computer systems' work for years to come: "The expanded use of information systems also raises major challenges to the traditional forms of administration and authority, the right to privacy, the nature and form of work, and the limits of calculative rationality in modern organizations and society."

Information Ecology: Open System Environment for Data, Memories and Knowing

- Sabine Grabner (MCR)

Baker, K.S. and G. Bowker. 2007. Information Ecology: Open System Environment for Data, Memories and Knowing. Journal of Intelligent Information Systems. BDEI Special Series. Online: http://www.springerlink.com/ content/1865762173811847/ fulltext.pdf

Literally citing the abstract of this paper, it gives insight in to the 'traditionally unreported invisible work' of knowledge management in the field of ecology. The expression that comes to my mind when recommending this article to the LTER IM community is 'preaching to the converted'. However, we keep improving our skills by exchanging our knowledge about knowledge, and this paper actually has quite an abstract take on it. That is why I recommend this paper by K. Baker and G. Bowker that summarizes their one year study on the heterogeneity of information and its management within the LTER community. Even more importantly, I hope this paper will find its way into the hands of LTER scientists to increase their temporal horizon from the lifetime of their own work to the lifetime of the ecosystem. This paper provides a comprehensive introduction to the vision, concepts, acceptance, applications, constraints, issues, and costs of contemporary knowledge management based on a real-world study. Understanding the concepts of knowledge management will help to understand why we (the Information Managers) spend so much of our time nagging scientists for things like metadata and protocols.

I liked the quote in the summary: 'An information environment isn't something you finish, it's something you start'. It is almost like saying an Information Manager is a new species with good chances of survival at times of increasing species extinction in our fast-forward world.

◆ Frequently Asked Questions

What is the rationale for publishing DataBits twice a year?

- Karen Baker (PAL && CCE) and John Campbell (HBR)

DataBits is an electronically distributed newsletter; hardcopies are often distributed at the LTER CC Meeting - a bit of reading for those quiet moments. Though there are historical ties to community meetings that have influenced the timing of DataBits publication, there is a contemporary update to this rationale. The IM Committee traditionally meets once each year in the summer. Until 2006, the LTER governance structure called for two science Coordinating Committee (CC) meetings a year: one in the Spring attended by two members from each site and one in the Fall attended by one site representative. DataBits, initially developed in 1990 (Porter, personal communication), was redesigned with a rotating editorship and a two issues per year publication cycle in 1999 (Baker and Brunt, DataBits Spring 1999). This timing allowed it to be distributed at each CC meeting. Thus the newsletter provided both an informal publication venue for the LTER Information Management (IM) committee participants and associates as well as a communication mechanism bridging the information management and site science communities.

The LTER by-laws were updated in 2006 (Zimmerman, Network News, Fall 2006 p5) in order to enhance governance of the larger community of sites. At this time, the CC was transformed and renamed the Science Council with one general in-person all sites meeting planned for Spring. Despite this change, publication of DataBits will continue twice a year: first in the Spring (coordinating with the Spring SC meeting) and once in the Fall (serving to keep the pace of communication at half-yearly and to serve as a forum for summarizing the annual IM meeting). A biannual publication cycle ensures continued timely communication of newsworthy IM topics among LTER information managers, scientists, and the broader ecological community.

Calendar

June 5-9, 2007 – 5th International Symposium on Digital Earth - San Fracisco, California (http://www.isde5.org/)

July 9-11, 2007 – 19th International Conference on Scientific and Statistical Database Management - Banff, Canada (http://ssdbm2007.cpsc.ucalgary.ca/)

July 23-27, 2007 – GEOWEB 2007 - Vancouver, British Colombia (http://www.geoweb.org/)

August 5-10, 2007 – Ecological Society of America/Society for Ecological Restoration International Joint Annual Meeting - San Jose, California (http://www.esa.org/sanjose/)

August 2-4, 2007 – Annual LTER Information Managers' Meeting - San Jose, California