LTER Databits
Information Management Newsletter of the Long Term Ecological Research Network

Fall 2011

In this issue of *Databits*, we see results from group and individual efforts to better serve the LTER Network, both from a technological and a socio-technical perspective. Noting that "one would be hard-pressed this fall to find an idle information manager," IMC Co-Chairs Don Henshaw and Margaret O'Brien provide a detailed synopsis of the activities that have taken place since our last issue. Among the few activities Don and Margaret don't explicitly mention are those of *Databits* co-editors Richard Cary and Theresa Valentine, whom I have the privilege of thanking for their help in editing, formatting and even contributing to this volume. I am also very grateful to everyone who contributed content.

Each article either directly addresses network level projects or demonstrates tools that would be useful in increasing connectivity both within and across sites. Two major gatherings of information managers prompted reports – the Environmental Information Manager's Conference (EIMC), in Santa Barbara, CA, and the SensorNIS Conference at Hubbard Brook, in New Hampshire. In addition, reports from smaller group efforts, including those focused on the Drupal Environmental Information Management System (DEIMS), LTERMapS, and the Metabase, demonstrate the progress these groups made by using common frameworks to increase their technical resource base and more transparently illustrate the linkages between research and data. At individual sites, we are introduced to resources other sites can leverage to improve the connectivity of their own data, including a process to smoothly retrieve national climate and streamflow data, the ability to contextualize photos with spatial information, and the means to generate and use web services to streamline the population of ClimDB and HydroDB.

This issue also provides two outside commentaries from colleagues who have a stake in the accessibility of the LTER Network. Prompted by his attendance of the EIMC, Dr. Irbis Gallegos, of University of Texas, El Paso's CyberSHARE Center, reflects on the overlap between his research and that of the LTER information management community. Gallegos discovers great potential for multi-disciplinary improvements in the management of ecological data through collaborations that extend beyond the LTER Network. Dr. Ted Gragson, Coweeta's Lead Principal Investigator, reflects on the importance of integrating site science and information management, noting that individuals, not technologies, are the most important drivers of integration and accessibility in a network such as ours.

As we continue conversations about what kind of network the LTER is and how it functions, it may be useful to examine a variety of analogical models. One excellent approach lies in the models presented by Albert Barábasi's book, *Linked: the new science of networks*. In it, Barábasi describes "scale-free networks" in which "hubs" naturally emerge to efficiently link the many nodes that develop in a context of self-organizing complexity. Since information managers frequently are asked to use many kinds of technology across a variety of domains and sites, we have the opportunity to act as at least one kind of hub to connect the many nodes of LTER. If the analogy proves applicable, then the articles here suggest a measure of success in doing that job well.

-- John F. Chamblee, CWT IM, Fall 2011 Databits Co-Editor

Featured Articles

Commentary

News Bits

Good Tools And Programs

Featured Articles

Mining Long-term Data from the Global Historical Climatology Network

edit

Wade Sheldon (GCE)

Introduction

Long-term climate data are critically important for climate change research, but are also needed to parameterize ecological models and provide context for interpreting research study findings. Consequently, climate data are among the most frequently-requested data products from LTER sites. This fact was a prime

motivating factor for development of the LTER ClimDB database from 1997 to 2002 (Henshaw et al., 2006). However, direct climate measurements made at the Georgia Coastal Ecosystems LTER site (GCE) are currently fairly limited, both geographically and temporally, because our monitoring program began in 2001. Therefore, in order to put results from GCE studies into broader historic and geographic context and to support LTER cross-site synthesis projects, we rely on climate data collected near the GCE domain from an array of long-term National Weather Service stations operated under the Cooperative Observer Program (NWS-COOP, http://www.nws.noaa.gov/om/coop/).

Data from NWS-COOP stations are distributed through the NOAA National Climatic Data Center (NCDC, http://www.ncdc.noaa.gov/oa/ncdc.html), so we have periodically requested data from NCDC for these ancillary weather stations to supplement GCE data. This process was greatly simplified in 2009 when we developed fully automated NCDC data mining support for the GCE Data Toolbox for MATLAB software (Sheldon, 2002). Functions in this software package proxy all interaction with the NCDC web site to retrieve and parse daily summary data for any NWS COOP station, then execute a workflow to convert English units to metric, perform QA/QC checks, and apply both documentation and attribute metadata to produce a fully documented tabular data set ready for analysis or synthesis. Unfortunately, this entire process ground to a halt in April 2011 when NOAA announced that it was abandoning the traditional COOP/Daily data forms, meaning that daily summary data sets would not be available from the existing web application beyond December 2010. We clearly needed to find a new source for NWS-COOP data.

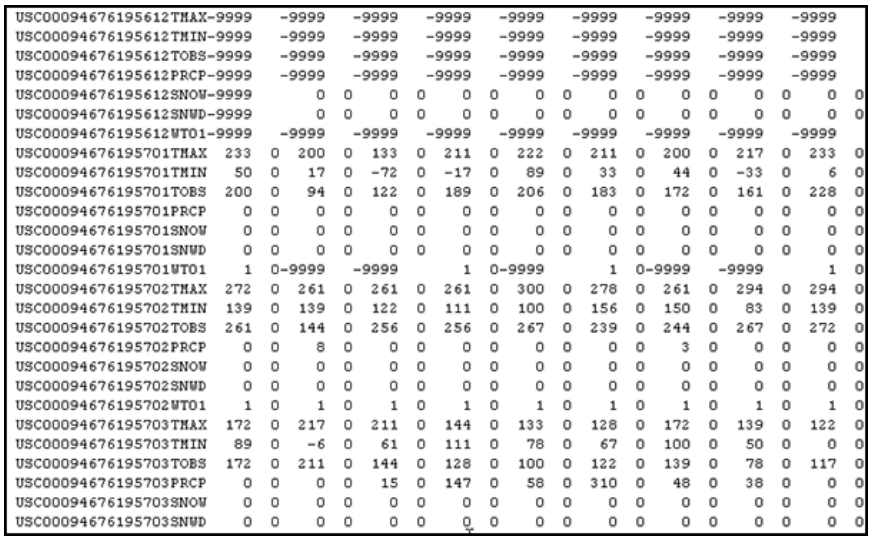
Goodbye COOP/Daily, Hello GHCN-D

When the NCDC announced the termination of the COOP/Daily forms (memo), they stated that these data products were being replaced with Global Historical Climatology Network Daily (GHCN-D) data that will be freely available in the future through a revised Climate Data Online (CDO) system. After seven months of limbo, the new NCDC CDO system was officially unveiled in November 2011 and is now available for use (http://www.ncdc.noaa.gov/cdo-web/search). The new system is a major improvement, both in terms of appearance and usability, but the revised web interface relies on client-side Javascript interaction for selecting stations and parameters and data download links are only exposed in email responses to data requests. This new architecture therefore precludes agent-based data mining using MATLAB or other software incapable of intercepting email messages. So in terms of restoring access to NWS-COOP data for GCE this new system is two steps forward, one step back.

Rather than returning to interactive web-based data requests for all of our stations, I searched for another way to access GHCN-D data that is more amenable to data mining. I discovered that GHCN-D data files are also available from NCDC via anonymous FTP, albeit in formats only a 1960's Fortran programmer could love (ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/). A single file containing the full period of record for each GHCN station (including all NWS-COOP stations) is available in the "all" subdirectory of the FTP site, as a text file with a ".dly" extension. For example, our NWS station on Sapelo Island (COOP ID 097808) is available as USC00097808.dly. Therefore just knowing the COOP station ID is sufficient to form an FTP URL programatically that will retrieve all available data for a station.

GHCN-D Format - The Bad, the Really Bad, and the Ugly

NOAA has never been known for distributing easy-to-parse data files, but the GHCN-D files set a new standard for user-unfriendliness (fig. 1). Each data row is preceded by a 21-digit label containing the station id, year, month, and a 4-digit parameter code - all concatenated without spaces. Following this label are 31 repeating groups of 8-digit fields, each containing a 4-digit signed integer value (yes, integer) and 3 separate qualifier codes, one group for each potential day in a calendar month. Months with less than 31 days are padded with -9999 missing value codes and no qualifiers. In order to store any type of numeric value in a 4-digit integer field, many parameters are represented in unusual units of measurement (e.g. tenths of a degree celsius, hundredths of a millimeter). In all, 189 distinct parameters are possible after compound, coded parameter types are expanded (e.g. SN*# is minimum soil temperature, where * is a 1 digit ground cover code and # is a 1 digit depth code). Working with these files clearly requires custom programming, advanced data integration methodology, and a good sense of humor.



USC00094676195612TMAX	-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999	
USC00094676195612TMIN	-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999	
USC00094676195612TOBS	-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999	
USC00094676195612PRCP	-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999	
USC00094676195612SNOW	-9999		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
USC00094676195612SNWD	-9999		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
USC00094676195612WT01	-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999		-9999	
USC00094676195701TMAX	233	0	200	0	133	0	211	0	222	0	211	0	200	0	217	0	233	0	217	0	233	0	217	0	233	0	217	0	233	0	217	0
USC00094676195701TMIN	50	0	17	0	-72	0	-17	0	89	0	33	0	44	0	-33	0	6	0	-33	0	6	0	-33	0	6	0	-33	0	6	0	-33	0
USC00094676195701TOBS	200	0	94	0	122	0	189	0	206	0	183	0	172	0	161	0	228	0	161	0	228	0	161	0	228	0	161	0	228	0	161	0
USC00094676195701PRCP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
USC00094676195701SNOW	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
USC00094676195701SNWD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
USC00094676195701WT01	1	0	-9999		-9999		1	0	-9999		1	0	-9999		-9999		1	0	-9999		-9999		1	0	-9999		-9999		1	0	-9999	
USC00094676195702TMAX	272	0	261	0	261	0	261	0	300	0	278	0	261	0	294	0	294	0	294	0	294	0	294	0	294	0	294	0	294	0	294	0
USC00094676195702TMIN	139	0	139	0	122	0	111	0	100	0	156	0	150	0	83	0	139	0	83	0	139	0	83	0	139	0	83	0	139	0	83	0
USC00094676195702TOBS	261	0	144	0	256	0	256	0	267	0	239	0	244	0	267	0	272	0	267	0	272	0	267	0	272	0	267	0	272	0	267	0
USC00094676195702PRCP	0	0	8	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
USC00094676195702SNOW	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
USC00094676195702SNWD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
USC00094676195702WT01	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
USC00094676195703TMAX	172	0	217	0	211	0	144	0	133	0	128	0	172	0	139	0	122	0	139	0	122	0	139	0	122	0	139	0	122	0	139	0
USC00094676195703TMIN	89	0	-6	0	61	0	111	0	78	0	67	0	100	0	50	0	0	0	50	0	0	0	50	0	0	0	50	0	0	0	50	0
USC00094676195703TOBS	172	0	211	0	144	0	128	0	100	0	122	0	139	0	78	0	117	0	78	0	117	0	78	0	117	0	78	0	117	0	78	0
USC00094676195703PRCP	0	0	0	0	15	0	147	0	58	0	310	0	48	0	38	0	0	0	38	0	0	0	38	0	0	0	38	0	0	0	38	0
USC00094676195703SNOW	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
USC00094676195703SNWD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

figure 1. Screen capture of a GHCN-D data file. A 21-digit label is followed by 31 8-digit fields, each containing an integer value and 3 qualifiers for the corresponding day of the month.

Developing a GHCN-D Parser

The first requirement for using this new data source was developing a MATLAB-based parser to support importing GHCN-D files into the GCE Data Toolbox. After trying several different approaches, the simplest strategy proved to be generating an intermediate text file that is refactored as a normalized data table. The normalized table contains separate columns for station, date (including day), parameter code, and 3 distinct flag columns (fig. 2). In addition to simplifying loading the file into MATLAB, this step provided an opportunity to filter out null values for invalid dates *a priori* (e.g. February 31) and add programmatic support for date range filtering to replace the lost functionality of the NCDC data request form (i.e. date range selector fields).

---

name:Station	Date	Parameter	Value	MFlag	QFlag	SFlag
units:none	MM/DD/YYYY	none	none	none	none	none
datatype:s	s	s	d	s	s	s
variabletype:nominal	datetime	nominal	data	code	code	code
numbertype:none	none	none	discrete	none	none	none
precision:0	0	0	0	0	0	0
USC00097808	05/01/1957	TMAX	244	~	~	0
USC00097808	05/02/1957	TMAX	283	~	~	0
USC00097808	05/03/1957	TMAX	261	~	~	0
USC00097808	05/04/1957	TMAX	267	~	~	0
USC00097808	05/05/1957	TMAX	194	~	~	0
USC00097808	05/06/1957	TMAX	233	~	~	0
USC00097808	05/07/1957	TMAX	239	~	~	0
USC00097808	05/08/1957	TMAX	244	~	~	0
USC00097808	05/09/1957	TMAX	256	~	~	0
USC00097808	05/10/1957	TMAX	256	~	~	0
USC00097808	05/11/1957	TMAX	267	~	~	0
USC00097808	05/12/1957	TMAX	267	~	~	0
USC00097808	05/13/1957	TMAX	272	~	~	0
USC00097808	05/14/1957	TMAX	272	~	~	0
USC00097808	05/15/1957	TMAX	283	~	~	0
USC00097808	05/16/1957	TMAX	278	~	~	0
USC00097808	05/17/1957	TMAX	278	~	~	0
USC00097808	05/18/1957	TMAX	272	~	~	0
USC00097808	05/19/1957	TMAX	300	~	~	0
USC00097808	05/20/1957	TMAX	328	~	~	0
USC00097808	05/21/1957	TMAX	317	~	~	0
USC00097808	05/22/1957	TMAX	294	~	~	0
USC00097808	05/23/1957	TMAX	294	~	~	0
USC00097808	05/24/1957	TMAX	289	~	~	0

figure 2. Screen capture of a normalized intermediate file generated by the MATLAB-based data parser. The header fields contain attribute metadata tokens that facilitate importing the file into the GCE Data Toolbox software, and tilde characters (~) are used to represent null strings to ensure efficient parsing.

After getting over the initial parsing hurdle, the next challenge was de-normalizing the derived table to generate a conventional ("wide") tabular data set and converting values into appropriate numeric formats and standard units. I began by building a table of all 189 potential parameters that could be present in a GHCN-D file, obtaining parameter codes, definitions and units from the "readme" file in the FTP directory. I then added columns for attribute metadata descriptors for the GCE Data Toolbox (data type, variable type, numeric type, and precision), along with original units, target units, and unit conversion multiplier (as appropriate) for each parameter. The last step was to import this table into the GCE Data Toolbox to serve as a reference data set for the import filter function.

To generate a conventional tabular data set, the intermediate data file is loaded and queried using the GCE Data Toolbox to generate a list of all unique parameter codes present. The import filter then serially subsets data records by parameter, looks up parameter characteristics from the reference data set, applies the attribute metadata descriptors (converting units as specified), and joins the data records together by date. The result is a derived tabular data set with columns for station id and date, and paired value and qualifier flag columns for each parameter in the original GHCN-D file (e.g. TMAX, Flag_TMAX, TMIN, Flag_TMIN, etc.). Values in these columns are converted to their appropriate data type (e.g. floating-point, exponential, integer) for the accompanying National Institute of Standards and Technology (NIST) SI units. To complete the GCE Data Toolbox data structure, the provider-defined qualifiers are converted to intrinsic QA/QC flags, meshing flags with any flags assigned by QA/QC rules defined in the metadata template and evaluted during processing.

This entire workflow is now encapsulated into a single import function in the GCE Data Toolbox software (imp_ncdc_ghcnd.m), allowing any GHCN-D file to be loaded and transformed in a single step. A second function (fetch_ncdc_ghcnd.m) was also developed to handle the FTP request and file download, then call the import function to parse the data. Together, these functions restore the capacity to mine data for any NCDC climate station over the Internet using the GCE Data Toolbox software (fig. 3). These new functions are available now on request and will be included in the next release of the GCE Data Toolbox software in December 2011 or January 2012. Copies of the parameter metadata table (in spreadsheet form) and examples of raw and normalized intermediate files are also available for those wishing to develop their own parsing solutions for this resource.

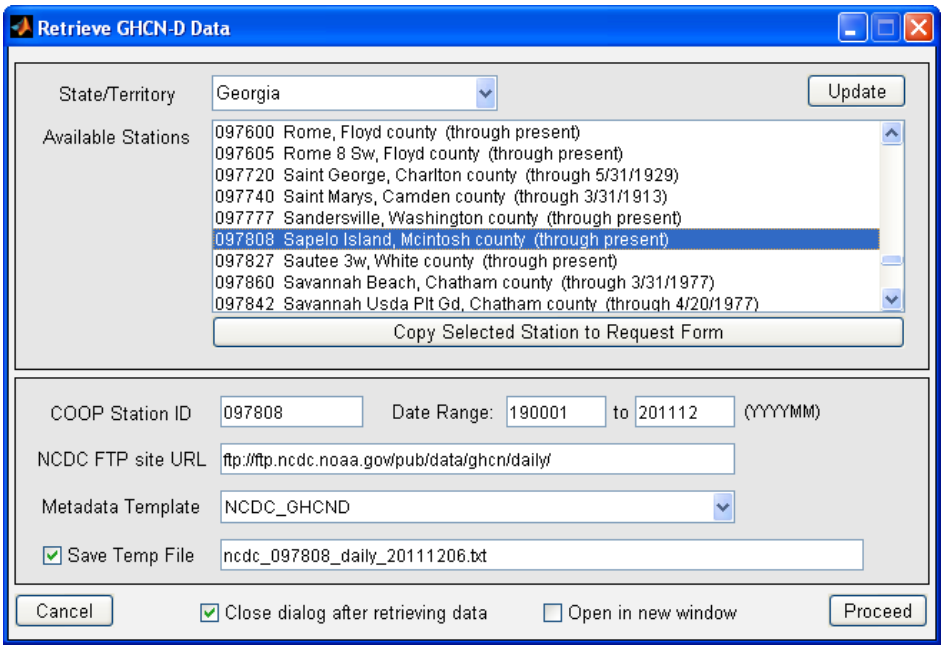


figure 3. Screen capture of the GCE Data Toolbox import form for retrieving GHCN-D data from the NOAA NCDC FTP site.

Conclusion

A wealth of environmental data are available on the Internet from U.S. and international monitoring programs. However, mining these data remains a significant challenge. Even when programs are written to automate this complex process, subtle changes to data formats, web interfaces, firewall rules and access protocols can render programs obsolete overnight. However, the efficiency gained by transitioning from human-mediated data mining to computer-mediated data mining can be tremendous, justifying this ongoing effort. LTER can also take a lesson from the new NCDC Climate Data Online web site: a slick web interface is a worth-while goal, but should not come at the expense of broad data access to the scientific community.

References

Henshaw, D.L., Sheldon, W.M., Remillard, S.M. and Kotwica, K. 2006. ClimDB/HydroDB: A web harvester and data warehouse approach to building a cross-site climate and hydrology database. Proceedings of the 7th International Conference on Hydroscience and Engineering (ICHE 2006). Michael Piasecki and College of Engineering, Drexel University, Philadelphia, USA. (http://hdl.handle.net/1860/1434)

Sheldon, W.M. 2002. GCE Data Toolbox for MATLAB. Georgia Coastal Ecosystems Long Term Ecological Research Program. (https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox/).

The Santa Barbara Coastal (SBC) LTER's implementation of projectDB using Metabase

edit

Margaret O'Brien (SBC)

This article provides an overview of our efforts to integrate a network-wide resource for describing research projects, Project DB, with our local solution for project data. ProjectDB was developed in a cross-site collaboration of LTER information managers to create software tools to track and catalog research projects (Walsh and Downing, 2008; O'Brien, 2009). Metabase is an extensive relational database in production at GCE LTER, and which was adopted by MCR and SBC in a PostgreSQL implementation. Together, these tools provide the potential to extensively document and present to the public linkages between broad research projects and themes and the data that support them. The remainder of this article describes the tools and tasks involved, outlines our implementation thus far, and discusses future plans, the lessons learned, and the advantages gained to this point.

The tools comprising ProjectDB are

1. an XML schema specification (lter-project-2.1.0), which is based on EML 2.1;

2. XSL stylesheets for HTML display of the XML content;

3. a Javascript for handling a tabbed layout;

4. Cascading Style Sheets for elements which benefit from uniform presentation;

5. XQuery files to return project XML from an eXist XML database.

The system is organized to be available from the Network SVN repository, and all files may be housed in and called from an eXist XML database. The tools work together, and as-is, can be used to display project XML documents from within HTML “frames”. However, SBC chose to integrate projectDB into its information system, as GCE did in 2010 (Sheldon, 2010).

Metabase already has components for housing all information needed for LTER-projects. One of our major goals for Metabase is to export EML datasets, and we chose to first export LTER-project XML. This was advantageous for several reasons: first, the LTER-project schema is a subset of EML schema, and so limiting the scope of exports represented a lower hurdle that would quickly demonstrate the usability of both Metabase and projectDB. Secondly, redesigning the research section of our website was a high priority for SBC scientists. Third, "projects" represent only a fraction of the features available in Metabase, and so made it a good entry point for someone new to the design. Lastly, other Metabase features and content needs can be planned based on lessons learned from these project exports.

We tackled the description of our research projects in phases, and for Phase I (conducted in 2011) limited the work to high-level research "themes" rather than including details of specific activities. The themes will provide groupings for descriptions of specific activities to be added later. We expect that in Phase II we will include specific core activites and add additional browse, search, and cross-link functionality. Phase III will include other non-core activities and associated collaborations.

Project requirements for Phase I were

1. allow for coordination and writing of SBC high-level research themes;

2. provide for information storage in Metabase with export as LTER-project XML;

3. display research themes on the web with a modification of the standard XSL to HTML transformation;

4. provide browsing of themes using SBC keywords groups for “Habitat” and “Core Research Area” in a manner similar to that used by the SBC data catalog so that from a user’s point of view, the two catalogs will have the same “look and feel.”

Specific tasks for Phase I were

1. define the high-level themes and collect information for personnel, abstracts, related papers, images, temporal coverage, keywords, and associated data. Themes did not include spatial coverage during Phase I;

2. a) store the information in SBC-Metabase, and b) export as LTER-project XML;

3. adapt the browsing and menu design used for SBC’s data catalog (O’Brien, 2010) for the project catalog;

4. adapt the presentation tools developed by the projectDB workshop (i.e., XSL stylesheets, Javascript and CSS) and in use at GCE (Sheldon, 2009);

5. plan for submission of SBC’s projectDB files to the Network SVN repository.

Task 1 (gathering information) required coordination over several months among most of SBC’s scientists and was conducted by SBC project manager, Jenny Dugan, and Lead PI, Dan Reed. Database content was prepared and uploaded to Metabase by the IM assistant, Alex Guerra (Task 2a). The information manager, Margaret O'Brien, conducted tasks 2b-5. Website-related tasks (3 and 4) required a total of about one month, including iterative design with several SBC scientists. Feedback and discussion from the MCR information manager, M. Gastil-Buhl, was invaluable throughout.

Implementation

Project XML from Metabase: Export from Metabase was performed with Perl, and script development was planned so that components can be reused for exporting EML datasets. For Phase I, we did not use an eXist database to hold and query XML; instead, XML was exported to the website's 'lib' directory, which was adequate for our current query needs. Storing static files also meant that output could be easily examined and checked for schema-validity with desktop tools (e.g., Oxygen), which is essential during script development.

Query and display projects: In Phase I, we planned to browse projects ('themes') using canned queries for "Habitat" or "Core Research Area." After making a selection, the user sees a list of projects each with a title, truncated abstract, and an image (Fig. 1). Since we did not implement XQuery for Phase I, we used XPath in the template, which was specifically directed to select certain <keywordSet> nodes (see below). This would not be efficient for a large number of projects, nor for complex queries.

The second view is of the project itself (Fig. 2). The script to return a single project was very simple, since Javascript controls the tabbed views. The script has only three lines of essential code as described earlier (O'Brien, 2009); i.e., calls to the XML content and the XSL stylesheet, followed by the transformation to HTML.

Figure 1 (left). Screenshot of the output for a query for SBC research themes related to "Population Studies". A short display of each project is provided, with links, and the menus appear again on the left.

Figure 2 (right). Default tab view used for a SBC research theme. Descriptions of detailed activities will use different or additional tabs, e.g., including coverage.

We made significant adaptations to the XSL stylesheet for project display, and SBC scientists provided considerable input on the presentation over several iterations. We also created wrapper stylesheets so that we can reuse components in later phases with more projects (e.g., "activities").

"Coverage" tab: For Phase I and research themes, we did not include the "Coverage" tab. Research themes were too broad to consistently report on sampling sites. We will add both temporal and geographic coverage to research activities in a later phase.

Images: Including images with a research description was very important to the site scientists as a way to engage interest. They felt that, alone, the project descriptions were somewhat dry, and while this was appropriate for information-packed dataset displays, descriptions of research should be more visually enticing. Images were added to the "Description" tab, and in other available blank space, e.g., on the "Personnel" tab. We were able to control the image that appeared by using the <associatedMaterial> element's 'category' attribute.

Related Data: We did not add a list of datasets to the research theme's "Products" mainly because these themes are broad and interdisciplinary, encompassing many types of data. Instead, we chose to highlight the data that could be associated with a particular theme. We added a tab for "Related Data", and duplicated the specific queries to the SBC data catalog that would display data collections of interest to someone browsing each theme. We used the <keywordSet> element for this purpose as well. Figure 3 shows the similarity between the "Related Data" tab in a project view (lower right) and the SBC Data Catalog index (upper left). In the project view, only a subset of choices is presented to the user, but otherwise the interfaces are nearly the same.

Figure 3 Comparison of SBC data catalog index and SBC Research Themes "Related Data" tab.

Project keywords were essential for both queries and for building links to related data, and for these, we used features of LTER-project XML that are not available in EML. Figure 4 shows the XML keywordSets for the project in Figure 2, above, that can be used independently. In EML, the <keywordSet> element has no attributes, but the LTER-project schema added an attribute called "name". So for returning *lists of projects* (e.g., as in Fig 1), SBC's code specifically selected only <keywordSet> nodes with a name attribute containing "query". To build the *forms for dataset links*, the code makes use of different <keywordSet> nodes, i.e., those with a name attribute containing "data".

Figure 4. Sample of LTER-project XML showing how keyword content can be used to drive different uses in the same project XML file. Note: at SBC, the Core Research Area Thesaurus contains additional terms and so the thesaurus name deliberately does not contain the string 'LTER'. However, since terms from the five LTER Core Research Areas are required for datasets, this is labeled as a distinct thesaurus.

Discussion

In future phases, SBC projectDB will incorporate descriptions of specific activities, such as sampling campaigns and student projects. We will also plan linkages from each activity to related research themes, and the products can be mores specific, e.g. to individual datasets.

For tabbed display, Javascript is faster and more straightforward than the "stage" method (with CGI parameters) that SBC's data catalog uses, which is not surprising, as Javascript works through the client rather than calling the server for each view. In each of its catalogs, SBC adapted existing code to save time, so we now can compare two different implementations of a similar process. When we next upgrade our data catalog, we will convert it to using the Javascript code for tabbed display to match our projectDB implementation.

In Phase I, SBC used simple XPath instead of true queries, which will not be practical when we have added activities, or need to build more complex cross-links. XQquery with eXist is one option for us, as the web services are already built. However, with Metabase we also have the option of building searches and cross-links within the database itself, and/or exporting projectXML directly as a web service. We look forward to exploring these options.

Building datasets from Metabase was efficient and straightforward. We look forward to further developing our XML export, both for projects and datasets with other features of PostgresQL. Our use of projectDB has highlighted some useful additions to the Metabase schema that we will communicate to GCE, CWT and MCR.

Visitors to the GCE and SBC websites can compare the two implementations (http:/gce.lternet.edu/public/research/projects.asp, http://sbc.lternet.edu/research/catalog). SBC has chosen to display only broad research themes at this time, and will add more detailed activities later. GCE described all its activities in their initial implementation. SBC made some significant changes to the XSL stylesheets, mostly to accommodate these broad themes. However the look and feel of the two displays is still quite similar and SBC "activities" (when added) will use a style almost identical to the GCE view. As projectDB develops as a network resource, more feedback is to be expected, and centralized storage of the most popular templates may make maintenance of a network look-and-feel almost trivial.

Our use of keywords in projectDB brought up some interesting points. As far as using the five Core Research Area (CRA) Keywords, we found that the process of attaching these terms to research themes was very straightforward, and all project themes 'fit' - often into only one CRA term. However, most sties (SBC included) seem to have difficulty when trying to attach the CRA keywords to datasets. This is most likely because data are described at a very granular level, whereas the CRA keywords are broad and thematic. But we found that using projectDB as a container for research descriptions was an effective way to link between the CRA term and data.

It was also very straightforward to use projectDB keywords to build queries to another catalog, but this would not have been possible without the "name" attribute for <keywordSet> nodes. The addition of this attribute has already been suggested as a useful enhancement to the EML schema as well. LTER's work with both EML and projectDB will demonstrate the advantages clearly.

References

O'Brien, M. 2010. Using EML in Your Local Data Catalog. LTER Databits, Fall 2010.

Sheldon, W. 2010. Getting started with eXist and XQuery. LTER Databits, Spring 2009.

Walsh, J. and Downing, J. 2008. ProjectDB – Planning and Development of a Collaborative Programming Effort. LTER Databits, Fall 2008

Commentary

A Progressive Story about the Integration of Information Management and Site Science

edit

Ted L Gragson (LPI, CWT)

This contribution to DataBits represents my reflections on the integration of Information Management and Site Science within the LTER. It is not, I stress, about the technical means for achieving integration. Rather it builds on the premise that Information Management and Site Science are both at their core about people. In Site Science, people are the ones who generate, test, and evaluate hypotheses in building and expanding theory. In Information Management, people help other people effectively use imperfect technology to achieve their objectives. It is not always clear we understand this truth in the LTER.

This article is a progressive story that involves real and fictional characters. These characters help me illustrate my point and why it matters as the LTER strives to integrate Information Management and Site Science. I don't pretend to offer a roadmap for how to achieve integration, but instead merely comment on practices in relation to success.

Steve Jobs, who it is fair to say was known around the world, died on October 5, 2011. Three days later, on October 8, Dennis Ritchie, who was more likely known by only a select few around the world, also died. The accomplishments of these two very different computer giants have had inordinate influence far beyond the LTER, but offer important insights to current Network undertakings. Let us first consider Ritchie.

At the end of the 1960s Ritchie developed the general-purpose programming language *C*. In creating *C*, he gave birth to the concept of open systems in which a program written on one platform could easily be transferred to another platform. In the long run, *C* made it possible to port the *UNIX* operating system, which Ritchie co-developed with Ken Thompson, to any kind of computer. Now *UNIX* undergirds the modern digital world. Among other things, it is the foundation for Linux, Mac-OS, the Internet and most mobile devices.

Ritchie's fundamental contribution, however, was to the human dimension of technology. *C* made it possible for a programmer to have nearly universal skills across technology platforms by learning one operating system, one set of tools, and one language. That you have the necessary skill to develop for and use any computer terminal or cell phone, irrespective of the manufacturer, is a direct consequence of Ritchie creating the *C* programming language. Ritchie's focus was on the person, not the technology as evident in his statement "It's not the actual programming that's interesting. It is what you can accomplish with the end results that are important" (*Investor's Business Daily* Jan 27, 2003).

Jobs is a person about whom individuals from nearly every professional or use sector imaginable have opinions as to his contributions. Most significant to this story, however, is the fact that Jobs was not an engineer and that he was obsessed with making advanced technology simple to use. Starting with the groundwork he and Steve Wosniak laid for Apple in the early 1970s, Jobs can be credited with transforming technology to adapt to the people who use it – rather than forcing people to adapt to the technology. There may be legitimate concerns about Apple availing itself of reasonable restrictive power and extending it to unreasonable levels (Orland 2010) just like there are concerns about certain advantages claimed for the development cycle of Open Source Software (Bezroukov 1999). The point though is that Jobs was a principal architect of the shift from personal to personalized computing, which is foundational to democratizing technology.

In a democracy, the electoral system should enable every adult citizen to have an equal say in the decisions that affect their lives. The vote, in combination with the structure of the executive and the vertical distribution of power, are the means for translating individual wishes into a collective choice. The challenge in all such situations is to achieve governmental coherence and stability, representational legitimacy, a capacity to manage conflict, and an overall system responsive to internal and external influence (Diamond & Plattner 2006).

The nature of democracy and the accomplishments of individuals such as Jobs and Ritchie matter as we move irrevocably toward what is described as the new vision for LTER research. This vision seeks to understand human-natural systems by inextricably linking people and technologies in the context of

cyberinfrastructure to allow collaborative activities and advance technological solutions (Brunt et al. 2007). The vision as expressed on paper is good, but the future is always unknown and we must retain the ability to make strategic adjustments in the short term that will help ensure we arrive in the future we are seeking, rather than one we did not anticipate.

Charlie, in *Flowers for Algernon* (1966), provides a warning. He increased his intelligence and understanding of the world through technology, but he did so at the expense of the relationships with the people around him. In the end, he regressed, alone, into mental disability.

The report by Atkins et al. (2003) is very clear about the vast research opportunities created by cyberinfrastructure. However, it also clearly lists many dangers including the failure to understand technological futures, the failure to recognize social and cultural barriers, the lack of appropriate organizational structures, and technological balkanization (rather than interoperability) among disciplines. Technology can be seductive, as illustrated by an image that circulated widely in the days leading up to the establishment of the National Environmental Observatory Network (NEON). Sitting cross-legged on the ground at the center of a heavily instrumented three-dimensional patch of land perhaps 100 m^3 in volume was a young woman holding a laptop who appears to be a Native American. It was never clear what message the image was meant to convey, but the woman was very alone with her technology. IM is about people because we must ultimately enable individual users to effectively use imperfect technology to achieve their objectives. To fail at this task is to fail at the integration of Information Management and Site Science.

Succeeding at this task begins by actively listening to what someone is saying or trying to convey - or actively observing someone peck their way through a workflow. Active listening requires nice, smart, adaptable people willing to enable users to reach their objectives. When we speak of "what scientists can do", "what technicians can do" or "what information managers can do" we are no longer listening, we are formulating a plan for connecting things to each other. In short, we are in the process of building the wiring diagram for a computer network. In *Green Eggs and Ham* (1960), Sam-I-Am convinces the unnamed character in Dr. Seuss' book to try what he has refused to try 72 times before. Unlike Ritchie and Jobs who focused on adapting the technology to the user, Sam-I-Am forces the user to adapt to the technology.

The LTER Network Office (LNO) gives the impression it views each LTER site as simply a node within the Network. In reality, each site is itself a network such that the collective of sites is more properly a federated network of sites. One could say the LTER is a "network of networks", but that term seems to be reserved for discussing how LTER relates to GLEON, NEON, CZO, etc. Whatever we may call the LTER, it is crucial that we truly understand the practices of Information Management and Site Science before normatively moving forward with a plan for integrating the two.

Baker and Chandler (2008) refer to the complex co-evolving system comprised of interdependent people, practices, values and technologies in a particular local environment as an "ecology of information." In control theory, a dynamic system is controllable if it can be driven from any initial state to any desired final state within finite time (Liu et al. 2011). Our ability to control a system is generally taken as ultimate proof that we understand it, as demonstrated when we use a phone, drive a car, or synchronize a communication network. Data pipelines in a linear, point-to-point computer network benefit from economies-of-scale that are not found in an ecology of information characterized by the complexities-of-scale associated with reconciling the heterogeneous data, collaborative effort, and disciplinary semantics needed to understand human-natural systems. While control theory is well developed mathematically, there is much we don't know about controlling complex natural, human, and socio-technical systems. This is true whether we are referring to those in the world we seek to understand or those we create and seek to direct to a final goal.

Sites must deal everyday with information access and availability issues - some are under autonomous control of sites (e.g., buy more computers, build a better interface) while others are a direct consequence of the institution in which they find themselves embedded (e.g., a university's broadband infrastructure or changing network security policies). Either way, sites alone and sometimes together contend with best serving and enabling all users to reach their objectives - scientists, technicians, students, administrators, or users from elsewhere. LNO from its vantage point at the center of this federation has the opportunity to assist with interoperability and by leadership and commitment provide an efficient computation and communication infrastructure that democratizes the capital produced by "doing science" across the federated network of sites.

However, in line with the views expressed by Ritchie, the solution is not the objective - it is what we can accomplish with the end results that are important. Sites understand how brittle monolithic solutions are. This is because data, workflows, and user experience are not abstract concepts that merely need to be wired together properly.

When we fail at the site level to break solutions into their component tiers, we risk perpetuating the view that Information Management is nothing more than good technical design and Science nothing more than good experimental design. All those activities lying between data and users subsumed under "workflows" require partnerships between communities that are information aware and cognizant of both the epistemological and ontological issues associated with interdisciplinary research (Baker and Chandler 2008). This requires information managers and scientists making and honoring commitments because they are not merely plumbers, they are partners in fulfilling site-level objectives as they relate to the Five NSF Evaluation Criteria they will be judged on at mid-term reviews and renewals.

Pogo the Possum once remarked, "We have met the enemy and he is us." In keeping with the history of Walter Kelly's character and the present story, Pogo meant that each individual is responsible for their involvement in democratizing the nation state they belong to or the LTER small-world in which they participate. Sam-I-Am offers one way forward and Pogo another. Sensitivity to the long-term depends on procedures, mechanisms, and strategies that are continually articulated, responsively designed, and thoughtfully deployed through recurrent balancing, aligning, and negotiation between all parties vested in the objective (Karasti et al. 2006). A futuristic vision of a perfect technical solution for integrating Information Management and Site Science is nothing compared to being able to trust the advice of information managers and site scientists, or be assured they will honor their commitments both in the short-term and the long-term. This is because at its core, Information Management and Site Science are about people, and technologies while necessary, will never be sufficient.

Atkins, Daniel E., Kelvin K. Droegemeier, Stuart I. Feldman, Hector Garcia-Molina, Michael L. Klein, David G. Messerschmitt, Paul Messina, Jeremiah P. Ostriker, and Margaret H. Wright. 2003. Revolutionalizing Science and Engineering Through Cyberinfrastructre: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Arlington VA: National Science Foundation.

Baker, Karen S. and Cynthia L. Chandler. (2008). Enabling long-term oceanographic research: Changing data practices, information management strategies and informatics. *Deep Sea Research II* 55: 2132-2142. (doi: 10.1016/j.dsr2.2008.05.009).

Berzoukov, Nikalai. 1999. A Second Look at the Cathedral and Bazaar. *First Monday* 4(12). URL: http://firstmonday.org/issues/issue4_12/bezroukov/index.html (Accessed December 7, 2011)

Coleman, Murray. 2003. Ritchie's Code? Improvement; Aim High: Unix co-creator's perseverance changed the language of computers. Investor's Business Daily January 27, A04.

Diamond, Larry and Marc F. Plattner. 2006. Introduction. In *Electoral Systems and Democracy*, Larry Diamond and Marc F. Plattner Eds., pp. ix-?. Baltimore: Johns Hopkins Press.

Karasti, Helena, Karen S. Baker, and Elija Halkola (2006). Enriching the notion of data curation in e-science: Data managing and information infrastructuring in the Long Term Ecological Research (LTER) Network. *Computer Supported Cooperative Work* Vol. 15, No. 4, pp. 321-358. (doi: 10.1007/s10606-006-9023-2)

Liu, Yang-Yu, Jean-Jacques Slotine, and Albert-László Barabási. 2011. Controllability of complex networks. *Nature* 473(May 12): 167-173. (doi:10.1038/nature10011)

Orland, Kyle. 2010. Opinion: The Tyranny Of Apple's App Store Review Guidelines. URL: http://www.gamasutra.com/view/news/30346/Opinion_The_Tyranny_Of_Apples_App_Store_Review_Guidelines.php (Accessed December 7, 2011)

An Incoming Computer Scientist's View on Unifying Standards and Procedures in the LTER Community

edit

Irbis M. Gallegos (Visiting Assistant Professsor, University of Texas at El Paso)

I recently attended the LTER EIM'11 Conference in Santa Barbara CA. Even though I am a computer scientist with multidisciplinary research experience on developing cyberinfrastructure for geoscientists and environmental scientists, I must admit that I was excited about attending the conference. I knew I was about to meet top researchers from the environmental science community and, not only that, I was about to present what I thought would be the silver bullet for providing data quality assurance for scientific sensor data.

During the poster session at the conference reception, as I started walking around, meeting great people, and learning about all the fascinating projects developed within the different LTER communities, I discovered other work that overlapped to a great extent with my research effort. At first I felt anxious, thinking that I had spent the last three years of my research career working on something that was already done, but then I realized that this wasn't necessarily true. After stopping for a moment and thinking about the situation, I realized that even though some of the work was similar to mine, my work was really the next step towards where the LTER data quality communities could be possibly moving in a near future. Also, the fact that my research was so similar to others, even though it had been conducted independently from LTER efforts, assured me that I was in fact making progress and going in the right direction inasmuch as my work could contribute to the LTER community's data quality goals.

After I had convinced myself that my work was indeed novel, I attended the conference's birds-of-feather session on "automating data processing and quality control using workflow software." I heard great talks about efforts to automate data quality processes through workflow systems, a 10-year research effort to automate data quality checking using off-the-shelf back-end statistical analysis software, and the most recent effort on detecting environmental events at near-real time. The last two were of particular interest to me because that is the topic of my dissertation work.

At that point, I asked myself, "How is it that I spent three years working on a cyberinfrastructure approach to improve data quality processes in scientific sensor data and did not know about these other efforts?" I knew I had done a literature review, that I had talked to my environmental scientist colleagues about current efforts, and that I had even browsed through the LTER community website as I was looking for job opportunities. In addition, it also seemed that most of the data quality research used expert processes similar to my work to detect the anomalies in the data. Yet the expert knowledge associated with such processes did not surface in my literature review. I realized that the LTER community has an opportunity to improve the ways in which emerging ideas are disseminated.

The remaining text is based on my personal experience as an outsider to the community. In my opinion, one of the main challenges for the LTER community is to determine how to establish the means to effectively communicate the accomplishments and research efforts of its members. For example, at my university, I collaborate with scientists conducting research at the Jornada-LTER site. However, I am unaware of other research efforts at this site. My only contact with the LTER community is through the scientists with whom I collaborate. Other efforts are hard to discover. As a result, it is possible that other research efforts on data quality are being conducted at the same research site without others being aware. Thus, scientists lose the opportunity to collaborate and share information.

In addition, it appears that the LTER sites are further subdivided into smaller, individual communities each with its own data procedures. For instance, my research work has a component that involves identifying data quality processes that are common to the scientific communities that collect data through sensors. As soon as I started working on this component, I recognized that even the communities sensing the same type of data, e.g., Eddy covariance data, had different data quality processes. In addition, the data analysis and data verification infrastructures are different. Thus, attaining inter-operability and sharing data procedures between sites can be extremely difficult.

Another challenge is how to document and share the data standards and processes that should be common for the different LTER research sites. For example, I came across a well known issue in the data quality community as I was working on my research. Other approaches face the same issue of having to adapt to the formats of different data loggers. Consequently, data files need to be parsed and analyzed. Such a challenge has been addressed through the development of different "parsers" that read and interpret the data files. However, such an approach is not scalable because a new parser has to be developed for every type of data file that needs to be verified. This challenge could be addressed if the community defines a data standard for such data files. The standard, if endorsed by the community, would provide the means for scientific instrumentation vendors to develop instruments that generate consistent data formats. A similar need is evident with the data quality standards, e.g., measurement units, that could facilitate specification of data properties of interest.

The good news is that the LTER community has been building the sense of community needed for scientists to have confidence in the work of other colleagues. I find it promising that contrary to the common belief that scientists are reluctant to share openly collected and derived datasets, I find that scientists are willing to share the processes and tools used to create datasets. However, the fact that different efforts, e.g., those related to data quality processes, overlap with each other (even though they are developed independently at different institutions) highlights the need to establish procedures and processes for documenting and sharing efforts within the community, as well as establishing standards.

News Bits

A busy fall season for information managers

edit

Don Henshaw (AND) and Margaret O'Brien (SBC)

Don and Margaret are Co-Chairs of the LTER Network Information Management Committee

One would be hard-pressed this fall to find an idle LTER Information Manager (IM). Whether attending workshops, participating on Tiger Teams or writing production-oriented workshop proposals, network-level IM activities have been pervasive and every LTER site has been involved. Starting with the annual LTER IM Committee (IMC) meeting, the action has been non-stop. We highlight IMC activities here.

IMC/EIMC Meetings:

The IMC kicked off its annual meeting with a field trip and evening beach party before settling in for a one-day meeting. The IMC debated and approved our Terms of Reference document, which now formally describes our governance and decision-making processes. Active IMC working groups shared their current status and planned activities. The meeting focused on issues regarding assessment and usability of EML data packages, integration of ClimDB and other synthesis databases into the NIS PASTA framework, and presentation and discussion of the workflow concept within the PASTA architecture.

The IMC meeting was followed by the Environmental Information Management Conference (EIMC, https://eim.ecoinformatics.org), which LTER co-hosts with the National Center for Ecological Analysis and Synthesis (NCEAS), the Data Observation Network for Earth (DataONE), and the University of New Mexico every three years. LTER IM Corinna Gries (NTL) and Matt Jones of NCEAS were the 2011 conference co-chairs. The IMC uses this meeting format to interact with a wide variety of interested groups and individuals. This year's meeting featured five "Birds of a Feather" sessions, which enabled discussion of specific topics in smaller groups, concurrent with speaker presentations of accepted papers. Four of these were led by LTER information managers and pertained directly to LTER's current information challenges, and allowed for input from the full spectrum of EIMC participants.

Fall workshops:

LTER Information Managers organized and/or participated in a series of informatics-based workshops. Most are described in Bob Waide's LTER Network News article (http://news.lternet.edu/Article2446.html) but a summary follows here:

- 13-14 Oct: Sherri Johnson (AND) organized the StreamChemDB workshop in Corvallis, Oregon, to demonstrate and gather feedback for the cross-site hydrochemical database. IMs and researchers from LTER, CUAHSI, CZO, USGS, NEON, and USFS participated.
- 25-27 Oct: Lindsey Rustad (HBR) and Don Henshaw (AND) coordinated the use of NSF supplemental funds for 28 LTER IMs and researchers to attend the SensorNIS Environmental Sensor Network Workshop at Hubbard Brook, New Hampshire. Lindsey, Don and Corinna Gries (NTL) served on the planning and organizational team for the workshop. See the related DataBits article in this edition.
- 1-3 Nov: Theresa Valentine (AND) coordinated the LTERMapS production workshop with members of the GIS Working Group in Albuquerque.
- 8-11 Nov: Iñigo San Gil (MCM) led the second training workshop for the Drupal Ecological Information Management System (DEIMS) group.
- 14-16 Nov: Corinna Gries (NTL) coordinated the "Preparing instructional materials for a course in LTER information management" workshop in Albuquerque. From Corinna, "this course will first be taught to LTER graduate students and interested faculty in 2012. Thereafter, an online course is envisioned while the developed materials will be available for use by anyone on the LTER IM website at http://im.lternet.edu/resources/training_material."
- Oct-Nov: Jamie Hollingsworth (BNZ) visited the LNO for an extended period to work on a new cartographic almanac.

Tiger Teams:

Tiger Teams allow IMs, researchers, and community members to contribute expertise and provide feedback to Network Information System (NIS) developers on the design and development of the NIS. Two new Tiger Teams kicked off in November:

- Metadata quality: consider completeness and quality of metadata elements in establishing requirements for acceptance or rating the quality of a data package
- NIS data portal: consider the user interface for accessing LTER data

Product-oriented proposals for NIS workshops:

As with previous years, the IMC members developed proposals for product-oriented workshops and IM compensation related to NIS development. These 2012 working groups will further our recent work with data package usability assessment tools, NIS database integration, and data discovery and delivery.

- Requirements for the EML Dataset Congruency Checker (ECC) (O'Brien and Downing)
- SiteDB restructuring and enhancement to support the integration of ClimDB/HydroDB and StreamChemDB into the NIS (Henshaw)
- Enhancing NIS Data Portal to match the Data Portal function to LTER community needs (Pockman, Servilla and Porter)
- PersonnelDB Web Services Development at CCE (Connors, Sheldon, Bohm and Chamblee)
- EML Diagnosis and Best Practice Implementation Mentor (Gastil-Buhl and Bohm)

Network-level activities:

- Site information managers provide leadership roles on network committees, such as John Porter (VCR) as NISAC co-chair and Emery Boose (HFR) as the IMC representative on the LTER Executive Board.
- Recent discussions among LTER scientists show that issues of data accessibility are of LTER-wide interest, and many IMs participated in a broadly attended conference call.
- IM-Exec (http://im.lternet.edu/home/imexec/) continues to guide IMC activities. The IM-Exec winter meeting is being planned with a projected agenda to 1) consider site progress and products necessary to complement planned PASTA development, 2012-2014, 2) review data access issues, and 3) plan the IMC annual meeting, which will be held as part of the LTER ASM 2012 in Estes Park.
- We welcome another informative issue of DataBits, with this issue edited by John Chamblee (CWT) and with assistance from Richard Cary (CWT) and Theresa Valentine (AND).

LTER members participate in the Environmental Sensor Network Workshop at Hubbard Brook

edit

Don Henshaw (AND), Corinna Gries (NTL) and Lindsey Rustad (HBR)

From October 25-27, 2011, 72 people participated in the Northeastern Ecosystem Research Cooperative (NERC) Environmental Sensor Network / LTER SensorNIS Workshop at Hubbard Brook Experimental Forest LTER in New Hampshire. The meeting agenda, participants and notes are posted online at http://im.lternet.edu/projects/SensorNIS. The workshop focused on the emerging roles of environmental sensor technology and management in providing high quality, near real-time data on the physical, chemical, and biological properties of ecosystems. The workshop featured plenary talks from NEON, CUAHSI, USFS, USGS, LTER, NCEAS, and DataTurbine participants. Talks also highlighted sensor network capabilities at several LTER and northeastern research sites. Twenty-eight LTER participants from 20 LTER sites joined over forty northeast research station representatives and Hubbard Brook researchers to create a diverse mix of researchers, graduate students, land managers, information managers, and field technicians.

A pre-workshop survey showed that there is no lack of important and pressing research questions that new and/or more sensors could answer at this point in time. However, cost, technical and software know-how were the top reasons limiting sensor deployment, especially in cases where managing the data deluge requires specialists and pricy hardware solutions at the home institution. Remote locations pose an especially difficult challenge in terms of data transmission. Although almost all participants are maintaining a standard weather station, only about 15 % of the respondents stated that they have sensor systems at their site that meets their needs. A wide variety of systems are employed at these research sites. These are used in response to the different research questions, environmental conditions, and options for data transmission. Campbell Scientific hardware and software are the most frequently used technologies.

Much of the workshop was devoted to breakout group discussions that considered software and technology for both managing the data and performing quality assurance and quality control (QA/QC) in a streaming data context. Plenary talks included topics on the NEON Information System, the DataTurbine initiative, extensions to the Kepler scientific workflow system to accommodate sensor networks, as well as the Georgia Coastal Ecosystem MATLAB-based Data Toolbox. Sensor site establishment, types of sensors and communication platforms, data collection issues, data processing middleware, data archiving tools, and data access options and trade-offs were also considered. The types of quality control procedures that can be used for streaming data were examined and the need for documenting data screening and flagging procedures was emphasized. We also explored the idea of Data Levels to differentiate raw data from quality assured data and gap-filled, modified or derived data. The results indicate the need for a community-developed knowledge-base or "best practices" document for sensor network establishment and management.

An outline for a best practices document was generated as a workshop product and published online (http://im.lternet.edu/resources/im_practices/sensor_data). The outline follows the sensor data life cycle and on the data collection side considers selecting sensors, building a sensor platform, and choosing a site. The cycle outline includes streaming data from field to lab and managing the data streams with middleware to apply QA/QC routines and archival storage. Data access is considered at several steps along the way with the different goals of monitoring sensor health versus use to answer scientific questions. The intent is that this best practices outline will be populated interactively with short one-page descriptions by community members, although the meeting organizers will likely target specific individuals to provide specific content as a means to assure completion of the document.

The best practices for sensor networks document will provide a quick resource for sites establishing a sensor network or implementing a sensor management system. A research paper describing QA/QC procedures for streaming sensor data was also outlined and the authors will pursue publishing this. Other “next steps” include preparation of an LTER training proposal such as “tools and training for sensor network establishment and management”, virtual water cooler discussions on sensor network topics, proposing a sensor workshop for LTER ASM 2012 at Estes Park in September, and planning a future DataBits edition devoted to sensor network experience articles from sites. Meeting organizers will take advantage of future workshop products as a means of continually updating the online best practices document.

The workshop was funded in part through National Science Foundation (NSF) LTER supplemental funding for information management. NSF encouraged LTER sites to use these funds for coordinated, cross-site projects, and funds were granted to LTER sites to participate in this workshop. Additional funding was provided by NERC to cover northeastern participants and Hubbard Brook meeting costs. Lindsey Rustad (HBR) of the USFS Northern Research Station, Don Henshaw (AND) of the USFS Pacific Northwest Research Station, Corinna Gries (NTL), Jamie Shanley (USGS), and Peter Murdock (USGS) comprised the planning and organizational team. Hubbard Brook research staff provided considerable meeting support including hosting a site field trip and preparing meeting and poster session space.

GIS Birds of a Feather Sessions at Environmental Information Management Meeting

edit

Theresa Valentine (AND)

There were two GIS related Birds of a Feather sessions held at the 2011 Environmental Information Management (EIM) meeting. These sessions were facilitated by members of the LTER GIS Working Group.

Seventeen people attended the session on Geospatial Data Management for Ecological Research Organizations. Outcomes of the session were to develop a Wiki to share information and to develop methods to export and serve spatial data without having to store static files. There was quite a bit of interest in Open Source GIS solutions. The LTER GIS Working Group agreed to set up the Wiki, and invite participants.

The group started by identifying challenges and opportunities that information managers encounter with increase in the demand and volume of geospatial data and with integrating these data with research data collected as part of field studies. Several challenges were identified including providing access and analysis tools for large datasets, visualization tools, dealing with the exponential growth of data, capacity and organization, and getting researchers to think spatially.

The group shared their experiences with spatial data management, and folks were looking for automated methods for converting data from different projections and formats. The sense was that people need some tools to make the export of data automated to avoid storing data in multiple formats.

The Internet Mapping Session discussed software solutions, ways to meet the expectations and needs of clients, metadata and data portals, performance tuning, geo-tagging other data, and citations for applications. Outcomes include building a Wiki to share links to applications, tools, and other information. Over 20 people attended the session.

There were several paths that individuals chose to implement their map services, including google products, open layers, ArcIMS, ArcServer, MapServer, and GeoServer. People tended to pick the platform that worked best for them, based on budgets, staff, expertise, and the requirements of specific applications. Many participants were working within a Drupal framework to build their applications.

There was a lively discussion on how developers measure what the users want. Ideas were feedback from users (formal and informal), developing pilots and budgeting for re-development to change according to needs, and having a feedback button on the application.

More detailed notes on the two sessions can be found here:

http://im.lternet.edu/sites/im.lternet.edu/files/GeodatamanagementBOF2011.pptx

http://im.lternet.edu/sites/im.lternet.edu/files/Internet_Mapping_BOF2011_final.doc

The Drupal Ecological Information Management System (DEIMS): recent progress and upcoming challenges for a grassroots project

edit

Inigo San Gil (MCM)

Through the use of Working Groups and Production Workshops, the LTER Information Management Committee has provided examples of cross-site collaborations by leveraging expertise and resources across sites to provide resources for the LTER Network. The Drupal Ecological Information Management System (DEIMS) Working Group is not only an example of how such resources are developed, but also a demonstration of how they can be deployed and used by ecological information managers beyond LTER. This article presents a historical outline of DEIMS’ formation, summarizes recent successes, and outlines challenges for the future. In sum, DEIMS is making inroads based on the same collaborative principles of the LTER Information Management group and the open source community.

The DEIMS group was formed in 2008 by a handful of LTER Information Managers, as well as participants from the Organization of Biological Field Stations and the Oak Ridge National Laboratory. The earliest DEIMS collaborators were Sevilleta (SEV) and Luquillo (LUQ). At present, there are nine sites participating in the working group. Apart from the aforementioned early adopters and the LTER Network Office, these are Arctic (ARC), Jornada (JRN), McMurdo Dry Valleys (MCM), North Temperate Lakes (NTL), Niwot Ridge (NWT), and Plum Island (PIE).

During the 2011 training workshop in Albuquerque (see the photo gallery here), DEIMS participants made significant progress. Most member sites now have developer instances, or *sand boxes* where they can break rules, twist information content, and try new ideas that may be either discarded or adopted by the group. DEIMS is now also capable of interoperating with products that are released by other LTER working groups, including LTERMaps, the controlled vocabulary, and the Unit Dictionary. Group members have also begun using services that are developed by members of the community external to LTER, such as the NBII, the Encyclopedia of Life, and the ITIS species taxonomy service and other miscellaneous services such as current weather conditions. Consuming such services teaches us lessons about what works and what does not, without the costs associated with development and implementation. Instead, we get time to see the products in action.

At this point, DEIMS looks to be approaching a critical transformation stage. It is at this stage, sometimes referred to as *crossing the chasm*, where the highest failure rates occur among business start-ups. At this point, we expect DEIMS to undergo a period of unpredictable growth during which it will face its real tests as it is stressed by the need to scale up and meet operational needs. New project testers will find a myriad of feature enhancements and bugs that make demands on both the initial group and the broader community.

However, we argue that DEIMS is well positioned to face challenges related to crossing the chasm. A primary advantage is that we can *off-shore* most of the development work, maintenance load, and scalability issues to the thousands of Drupal contributors that lend their time free of charge to the community. DEIMS

uses the Drupal core and over fifty community contributed projects. In reality, the DEIMS team comprises of the direct work of hundreds of people across the world. Thanks to those unnamed contributors, DEIMS uses web services, list look ups, databases, relationships, controlled vocabularies, map applications, and excellent bibliographic management systems that were available “out of the box” for free.

In the end, the key to DEIMS’ success hinges on Drupal, with its large developer base and architectural flexibility. The Drupal API makes it appealing to both the skilled developer and the web aficionado. This structure, where all new projects hook into and leverage the core (as opposed to existing as monolithic addendums), gives further cohesion to the software and knowledge base of our group.

To read more about DEIMS, please see these peer reviewed papers

1. Case Studies of Ecological Integrative Information Systems: The Luquillo and Sevilleta Information Management Systems. Communications in Computer and Information Science, 2010, Volume 108, 18-35, DOI: 10.1007/978-3-642-16552-8_3.

   Resource at:http://www.springerlink.com/content/j183x10588574846/

1. Metadata Activities in Biology - Journal of Library Metadata, Volume 10, Issue 2 & 3 April 2010 , pages 99 - 118. Resource at:

   http://www.informaworld.com/smpp/content~content=a928352203~db=all~jumptype=rss

The DEIMS Google code project link is http://code.google.com/p/deims/. This project page contains also a link to the group mailing list, also hosted in Google groups.

Good Tools And Programs

Web Services for ClimDB/HydroDB Database

edit

Suzanne Remillard (AND) and Kyle Kotwica

In the early years of the ClimDB/HydroDB database development, experimental web services were developed to access the database and to demonstrate potential capabilities and its interoperability with harvest scripts and other software. These scripts have never been developed further, but they are operational. This article is intended to provide an introduction to their capabilities, as well as a practical example of their use.

The application, dubbed the Web Service Administration Module for B2B Application Methods (WAMBAM), is a Perl tool developed to facilitate web service creation, providing a Graphical User Interface (GUI) that ClimDB/HydroDB Administrators can use to create, edit, and delete services. WAMBAM creates an XML configuration file that contains the necessary information to build any ClimDB service. These services can be one of three types; SQL, Perl or JAVA. In an SQL service, the configuration file lists the connection information, name of the method, input and output variables, and the query string to access the data in the database. In the case of Perl or JAVA, similar information is provided, but instead of SQL, executable code is submitted either inline or by submitting a path to the code.

Services created by WAMBAM are accessible through a SOAP Server, in this case a CGI application that listens for WAMBAM web service requests. This application also serves as an outwardly facing web page that provides documentation, an actual Web Service Documentation Language (WDSL) file, and a sample client for each of the services offered. The WSDL is an Extensible Markup Language (XML) file that provides all of the information necessary to create a client and to use the services. The sample clients created are really just links to publicly available third party software that inspects the WSDL and creates a web page that allows users to access or test the services through a browser interface. In practice, it is expected that end users would create clients that access the services.

The current web service examples can be viewed and tested here: http://climhy.lternet.edu/wambam/soap_server.pl

When a request is made to a service, either through these generic browser clients or from a client on another machine, the request uses a handler (e.g., <service>_handler.pl) to receive the request and generate a response. Both the WSDL and the handler are virtual and are built on the fly with each call. The Soap Server listens for such requests.

Currently, two services have been built to access the ClimDB/HydroDB database. These are (1) climdb_raw, which accesses the raw data table, and (2) climdb_agg, which accesses the aggregated data table. Documentation can be found at http://climhy.lternet.edu/climhy_ws_api.html

Climdb_raw currently includes the following methods:

get_day - returns all the data for a given research site id on a given day;
last_harvest - returns the date of the last harvest for a given a site, station, and a ClimDB long or short variable name;
get_variable - returns the value for a variable between a given date range.

Climdb_agg service currently includes the following methods:

get_agg_monthly - returns the aggregated data from climdb_raw for a given range of years;
get_agg_yearly - returns the yearly value of the aggregated variable over a given range of years;
agg_over_all - returns the monthly value of the aggregated variable over the entire period of record.

At the Andrews LTER, we incorporate the use of these web services as a way of streamlining ClimDB/HydroDB harvests. This approach, detailed below, is an example of how these ClimDB/HydroDB web services can be used today. To update the Andrews ClimDB/HydroDB data, we use an automated process that requests the last date of data in the database for a given station and variable via the last_harvest methods of climdb_raw service. The following example shows our client, which is wrapped in the subroutine ‘get_start’ called with a site, station, and variable combination that we use in our harvest application to make a call to the climdb_raw web service:

```
$start = get_start('AND',$station,$name);
sub get_start {
        my $start = SOAP::Lite
        ->uri("urn:climdb_raw")
        ->proxy("http://climhy.lternet.edu/wambam/services/climdb_raw_handler.pl")
        ->last_harvest(@_)
        ->valueof('//date');
        }
```

The client includes the web service name, the handler (one could use the WSDL here), the method, and the return variable(s). Although the handler is built on the fly whenever the WSDL is inspected and will reflect any changes made to the service after it is deployed, the handler code is left on the server. Therefore it is

possible to bypass the WSDL and call the service directly. This is the approach we use in our application. The value returned from this request is obtained from ClimDB/HydroDB database using the last_harvest method and then is used as a start date to build our ClimDB/HydroDB exchange file. Our exchange file contains a separate header line for each site, station, variable combination. Although it is easy enough to create and harvest a large exchange file with all the station data in it, this web service enables us to easily customize each data harvest with only the newest data and creates a much smaller file that is more efficient to harvest into the ClimDB/HydroDB database.

Thanks to Don Henshaw and John Chamblee for editorial comments on this article

Automated Photo Geotagging

edit

John Porter (VCR)

The value of ecological photo archives is greatly enhanced when the location from which photos were taken, as well as the date and time, are known. Now for less than $100 you can purchase a "GPS Photo Tagger" that will allow you to capture and incorporate into the EXIF headers of your digital photos the locations where the photos were taken. No physical connection between the camera and "GPS Photo Tagger" is required. Timestamps are used to match up the GPS locations with individual photos, so any camera that records the date and time in an image can be used.

There are a huge number of "GPS Photo Tagger" models, varying in size, power requirements, interfaces for digital media (e.g., SD cards), and even interconnection to cameras. The things they have in common are

1. a built-in GPS unit;
2. frequent position logging (typically every 15 seconds) accompanied by sufficient memory to record hours to days worth of locations;
3. long battery life for continuous use - typically in the range of 8 to 24 hours;
4. time-stamped positions for use in relating position to photos taken at the same time;
5. USB connections for data download;
6. included software that edit existing .jpg image files to add geographic coordinates to the EXIF information in the file.

Things that tend to vary between units are

1. size - varying between a keyring fob (small) to a pager or cell phone (large);
2. display - some use only flashing lights to indicate proper operation, others have display screens for location and satellite information;
3. controls - some have only an on-off switch, others have multiple buttons and menu-based systems;
4. interfaces - some units depend on software running on a computer to do the photo geotagging, others have sockets for SD cards and other media and can do the geotagging using the unit alone. Some units have vendor-specific connectors that allow them to physically connect to a particular brand of camera equipped with a GPS interface or Bluetooth connection.

Another thing that varies, though not terribly much, is cost. GPS Photo Taggers range in price from about $50 to $200, with most in the $100-$150 range. Costs are a function of size (smaller=more), battery type & life (internal lithium vs. external, internal=more), amount of memory, interface (USB only vs. slot for SD card), software, display capabilities, motion detection (only record data when moving) and whether it has specialized support for particular brands of camera that allow it to do the geotagging internal to the camera.

An advantage of using these units is that the location information is inseparably encoded into the images themselves, so that even years later the location will be discernable. There are also a many tools, both on a local computer (e.g., Picasa) or on the web (e.g., Flickr or Panoramio) that allow the display of photo locations either on a map or in tools like Google Earth.