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

Spring 2013

It's a busy time for LTER information managers, as the 21 articles from 27 authors in this Spring 2013 issue of Databits can attest. We have an astounding number of workshop reports and announcements and network working group project updates this time, running from information management systems to sensor networks to GIS to NIS components. We also have a wide range of commentaries, good reads and tools, and articles from guest contributors to keep you glued to your screen, reading long past your bedtime. 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 is supported by rotating co-editors.

Aaron Stephenson (NTL) and **Hope Humphries (NWT), Spring 2013 Co-Editors**

Featured Articles

Commentary

News Bits

Good Tools And Programs

Good Reads

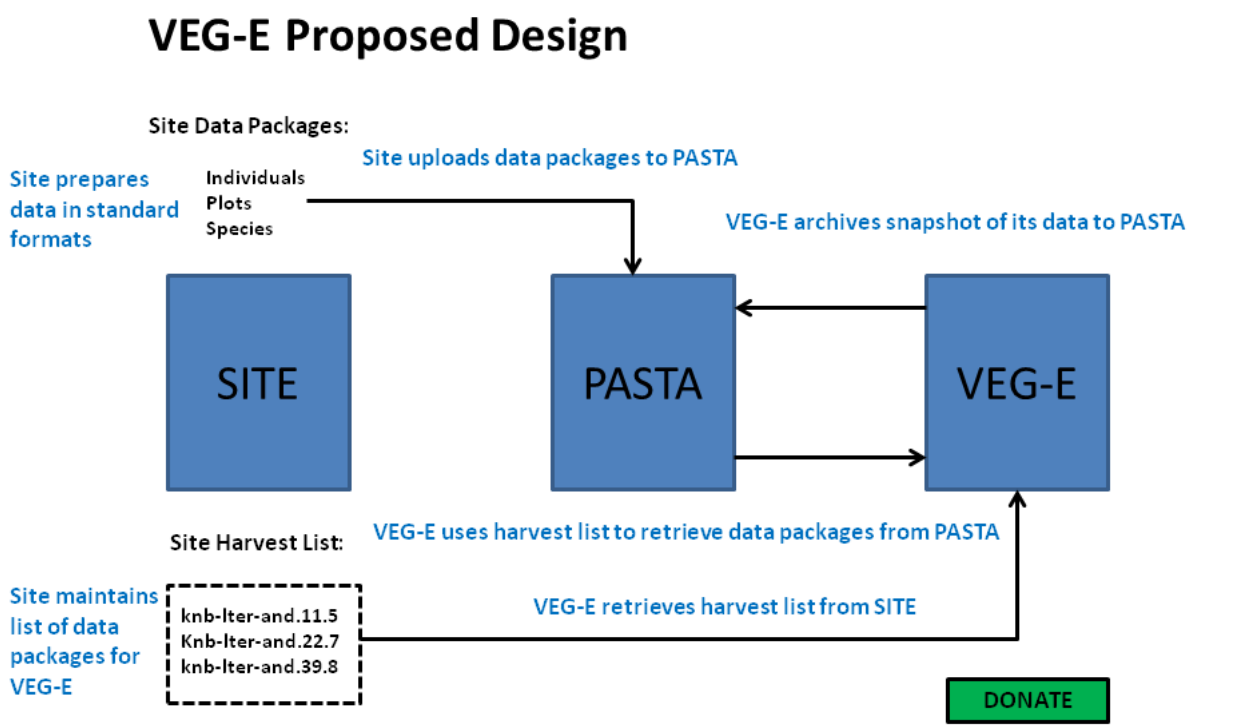
Featured Articles

A New Way to Use PASTA for Synthesis: Results from the Second VEG-DB Workshop

edit

Emery Boose (HFR), Fox Peterson (AND), Suzanne Remillard (AND), Don Henshaw (AND), Mark Harmon (AND)

At the recent VEG-DB workshop at the Sevilleta National Wildlife Refuge, April 30 to May 2, 2013, we addressed the question of how the new PASTA infrastructure might be utilized to support synthetic studies of long-term LTER vegetation data. These data have the potential to shed light on some interesting scientific questions, including: How does NPP depend on water potential? What is the role of mortality in NPP? How is productivity related to disturbance? Are scaling relationships between biomass and density constant across different taxa? The task of assembling and synthesizing LTER vegetation data from individual websites in order to answer questions like these is formidable and beyond the reach of most individuals or working groups. But a synthesis engine (VEG-E) that builds on the site and network investment in PASTA might bring the analysis of such questions within reach (see Figure).



In the workshop we identified two technical challenges to using PASTA for this purpose. The first is the heterogeneity of site data, which more often than not vary in critical details (structure, variable names, units) between sites and even within a site. This is especially true for biological data. For example, at SBC, growth in giant kelp (*Macrocystis*) is measured in g/m2-day and biomass is calculated as a function of height and frond number. At PIE, growth of marsh grass (*Spartina*) is measured by the change in blade number across seasons and biomass is calculated as a function of percent cover and blade number. And at AND, the biomass of Douglas Fir (*Pseudotsuga*) and other tree species is calculated as a function of DBH using various allometric equations, while ANPP is calculated as the change in biomass plus mortality.

Our solution to this problem requires participating sites to prepare their data in one of several prescribed formats, depending on the biome and the level of measurement (e.g. individual plant, plot, species), and to include critical derived variables such as biomass. This represents extra work for the site. However it greatly simplifies downstream processing, leading to a simpler and faster design for VEG-E.

The second challenge has to do with PASTA itself. As PASTA becomes fully populated it will contain a wealth of long-term vegetation data. However identifying the best datasets for synthesis will be non-trivial. For example, some sites (lumpers) may submit new data and corrections as updates of the same data package; while other sites (splitters) may submit new data as different data packages. Sites may have multiple vegetation studies and may have preferences about which data packages to contribute to VEG-E. Sites may also change their minds over time.

Our solution to this problem requires participating sites to post a harvest list of the datasets in PASTA that they would like to contribute to VEG-E. This list would be harvested at regular intervals (perhaps monthly) and used by VEG-E as a guide for which datasets to retrieve from PASTA. The list would be updated by the site whenever a new data package for VEG-E is submitted to PASTA or the site decides to change which data packages to submit to VEG-E.

The design of VEG-E itself was not considered in detail at the workshop. However it might be fairly simple and might consist (for example) of a backend relational database and frontend user interface. Though the number of individual records will be large, the data are well defined, the number of tables is limited, and many of the desired operations (e.g. aggregation and subsetting) are performed quite efficiently by database software. Wherever possible we would propose to use tools developed by others; e.g. for taxonomic reconciliation or graphing. In addition to a harvest process for retrieving site harvest lists and PASTA data packages, VEG-E would also include an archiving process for submitting snapshots of itself back to PASTA, perhaps on an annual basis, as a long-term record of VEG-E activities. Over time VEG-E could be extended to retrieve related data from other sources (e.g. climate data from ClimDB or plot information from SiteDB) using web services.

The VEG-E interface would include built-in tools for aggregating, subsetting, downloading, and graphing data and for generating simple statistical measures. Individuals could design their own workflows to download and analyze data from VEG-E, and users who prefer to do their own analyses could download the entire contents of VEG-E. By providing one-stop shopping for long-term LTER vegetation data, VEG-E would enable researchers to see at a glance what others are doing across the network, provide impetus for participating sites to keep their vegetation data complete and up-to-date, and provide a platform for synthesis of LTER vegetation data for the entire scientific community.

Though the focus of our workshop was on vegetation data, we believe the VEG-E model is inherently generic and could be adapted easily for other areas of interest, including climate, hydrology, stream chemistry, and soils. It might even be possible to design a single application engine that would serve multiple disciplines, utilizing PASTA and some additional data preparation at the site level.

Next steps for the project include working out details of the VEG-E design, specifying the structure (variables, variable names, and units) for site data for different biomes and different levels of measurement (e.g. individual plants, plots, species), preparation and submission to PASTA of data from 8 or 9 sites who will serve as early adopters, and development of a working prototype. Not to mention looking for funding.

Coweeta LTER Upgrades Sensor Stations by Implementing the GCE Data Toolbox for Matlab to Stream Data

edit

Richard Cary (CWT), John F. Chamblee (CWT)

Over the last 18 months the Coweeta LTER Information Management Team has been working to upgrade our sensor network so that we can stream the majority of our sensor data and publish to the web in near-real time. This article explains our motivations for pursuing this course, outlines our overall strategy for securing funding and planning the upgrade, and then outlines the processes and tools we have used in implementation. Our approach is characterized by the use of off-the-shelf software from sensor manufacturers (Coweeta standardized on Campbell Scientific sensors decades ago) and use of the GCE Data Toolbox for Matlab, available at https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox. We believe that by focusing on the adoption and use of existing, well-tested products, we have significantly reduced the time and financial resources necessary to undertake an upgrade and expansion of our sensor network.

Background

The Coweeta LTER (CWT) currently operates ten sensor-based field stations within the Coweeta Hydrologic Laboratory, near Otto, North Carolina. These stations measure a variety of parameters, including soil moisture, soil and air temperature, and photosynthetically active radiation. Historically, these sites have been based on Campbell Scientific CR10X dataloggers, and technicians would make monthly visits to each site to manually download the data. This method of data acquisition has a number of drawbacks, including high labor costs, diminished data quality due to undetected station down-time, and slow data publication rates.

The intensive labor costs associated with manual collection are tied to the time a technician must take to drive and then hike to each site. Not only does this effort consume time that could be spent on other projects, it also limits the number of sites that can be in operation at one time. A labor-intensive manual collection process also affects data quality. Since sites are only checked once a month, there is no way to know the status of the site until the monthly check occurs. This can lead to significant downtime due to any number of issues such as battery failure, sensor failure, vandalism, and interference from an animal or falling vegetation. These factors can result in the loss of up to one month of data.

Finally, a manual collection process implies, at least in the case of Coweeta, a manual data post-processing system for aggregating data and conducting quality checks. When accompanied by the slow rate of collection, these manual data aggregation and quality check processes reduce the frequency with which data are published – delays that can, in turn, delay research.

Pilot Project to Upgrade Equipment and Grant to Upgrade and Expand our Regional Network

To address these issues CWT information managers, technicians, and our Lead PI, Ted Gragson, submitted and received funding for two proposals to implement a near-real time data streaming system for all our field stations. Our pilot project was funded a by a $20,000 equipment supplement from the NSF and focused on the deployment of three stations, each using a different approach to data transmission and capture. Our pilot project was based on implementing workflows that were already in place at the GCE LTER. Based on our understanding of their work and the preliminary planning from the early stages of the pilot project, we also developed a plan to upgrade our entire sensor network and expand that network to encompass a greater proportion of CWT's regional study area. We formalized this plan into an NSF Field Station and Marine Lab (FSML) grant proposal that we submitted in January of 2012.

We began streaming data at our first pilot site in June 2012. At each station, we replaced the CR10 Dataloggers with CR1000 models, equipped them with wireless communication equipment, and implemented fully automated data processing, QA/QC, and publishing data products to the web using the GCE Data Toolbox for Matlab. We received word that the FSML grant was funded in August 2012.

The pilot project was successful in substantially increasing the frequency and quality of updates to data from these two sites, while at the same time significantly reducing labor costs for processing the data. Thanks to additional funding from the FSML grant, we have, as of this writing, upgraded the eight remaining stations at the Coweeta Hydrologic Laboratory and installed 3 new stations at Mars Hill College in Mars Hill, North Carolina. By the end of the year, we will have a network of fifteen sites in operation and we anticipate operating approximately 30 fully automated stations using these methods by August 2014. The remainder of this article provides details on the opportunities provided to us by the GCE Data Toolbox and the off-the-shelf software available from Campbell Scientific.

Data Transmission using LoggerNet

The Coweeta LTER is located in the southern Appalachian Mountains, and has very dense vegetation cover, along with steep, mountainous terrain. This terrain plays a large part in determining the wireless communication method that will be employed at each site, with 900mhz radios being the ideal option within the Coweeta Hydrologic Laboratory, where line of site can be established. Cellular modems are our preferred option outside of Coweeta, where cellular service is available. If neither option is available, we plan to use the GOES (Geostationary Operational Environmental Satellites) transmission system.

We use the Campbell Scientific LoggerNet software to manage most of the remote portions of the sensor network. LoggerNet handles scheduled data retrieval from the remote stations,includes an editor for creating datalogger programs, and can also be used to perform real time station status checks and update datalogger programs remotely if needed. Other solutions are available, but we've found this off-the-shelf software to be intuitive and user friendly and we believe it has helped us get this project off the ground more quickly than we might have otherwise.

Each cellular transmission station upgrade begins with an upgrade of the datalogger to a Campbell Scientific CR1000 model and the installation of a cellular modem with a Yagi directional antenna to transmit the data. Data transmission and timing are handled through LoggerNet, and data are downloaded every 24 hours to an offsite workstation at the University of Georgia. Our pilot site is located at a high elevation near the top of a mountain and is also under dense canopy cover. Normally, cell coverage is virtually nonexistent within the Coweeta Hydrologic Laboratory. However, we learned that at some high elevation sites, the cellular signal from the opposite side of the mountain is strong enough to allow us to establish a connection sufficient to transmit the relatively small amount of data involved. Each radio transmission site upgrade, just like the cellular transmission site upgrade, includes the installation of a CR1000 datalogger. Radio transmitter sites use either a 900mhz radio with an omni-directional antenna or a Yagi directional antenna, depending on if the site is a radio repeater station or not. In addition, a radio base station was established inside one of the on-site CWT LTER offices at the Coweeta Hydrologic Laboratory. The base station consists of a radio connected to Yagi antenna directed to the repeater station, and an Ethernet adapter. Connections to the radio network can be made through the LoggerNet software, which is also configured for daily data retrieval.

We have conducted one test of GOES satellite data streaming using an existing sensor array operated by the US Forest Service at the Coweeta Hydrologic Laboratory. Streaming data through the GOES system is more challenging for a variety of logistical and technical reasons. GOES users must have a federal sponsor (graciously provided, in our case, by our partners at the Coweeta Hydrologic Laboratory). In addition, while it is possible to download data using the natively supported GOES LRGS (Local Readout Ground Station) software, configuring this software for use with an automated workflow proved to be difficult. Instead, we contacted a local National Weather Service (NWS) office and requested that our sensor of interest be included in the Hydrometeorological Automated Data System (HADS) system. This allowed us to use the HADS support in the GCE data toolbox, but it required a data documentation process managed through intensive communication with the local NWS office.

NWS personnel are highly professional and wonderful to work with, but there is an additional investment of set up time for using the GOES system, when compared to the cellular or radio options. It is for this reason, as well as the fact that GOES bandwidth is limited nationwide, and that the communication is only one way, that we recommend GOES transmission only when directly managed options are unavailable.

Data Processing and Publishing with GCE Data Toolbox for Matlab

Once data are made available on a local computer by LoggerNet, the focus shifts to configuring the GCE Data Toolbox for Matlab. This software was developed by Wade Sheldon at the Georgia Coastal Ecosystem LTER (https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox) as a way to manage data by providing tools to perform QA/QC checks and error flagging, metadata creation and management, data transformation, and the creation and publication of data products. These tasks can be fully automated, making the GCE Toolbox an ideal solution for data streaming. While the Toolbox is available free of charge, it operates within the Matlab environment and requires a Matlab license.

Configuring the toolbox for the data streaming was fairly straightforward for two reasons:

1. The toolbox comes with a set of Demonstration products (located in the "Demo" folder) that are built for data streaming. The demo products can serve as a tutorial so that users can do a "dry run" of streaming, processing, and publishing a sample data set. However, they are also built to be copied into the userdata folder of the Toolbox (where the Toolbox can access custom and site-based projects) so that they can be modified and expanded to meet local needs.

2. Primary features of the Toolbox are custom import filters and metadata templates that can be written, stored, manipulated, and copied using the Toolbox's native interface.

We could have begun by using the Toolbox's standard text import filter to import the raw delimited ASCII data files into the toolbox. However, an important part of the hardware upgrade was the switch to CR1000 data loggers, which uses a Campbell TOA5 file format as a standard. Since one of the products available in the Toolbox's Demo folder is a TOA5 import filter, this format is easy to use with the GCE Toolbox's standard import filter. Once LoggerNet retrieved a file for us to work with, it took only a couple of minutes to pull in the file.

Once we had a standard GCE Toolbox Matlab file, we could enter dataset and variable metadata, as well as flagging criteria. When all of these metadata are entered they can saved as a metadata template and stored for re-application as part of a data harvest workflow. GCE Toolbox-driven data harvest, post-processing, storage, and publication is handled by three main GCE Toolbox functions that are in turn supported by a wide array of additional files.

The data harvester (data_harvester.m) is a generic workflow program that will receive arguments concerning source file path, import filter, metadata template, and publication file paths for a given data set; use parameters provided in other supporting scripts to retrieve the data from the data source, apply import filters and metadata templates; and then publish the data to the required location. Users can save copies of this program and modify them to meet their own needs, adding additional workflow items that are not stored in other stored components of the data harvest system.

The harvest timer (harvest_timers.mat) is a GCE Data Toolbox data structure that stores information on the arguments used to run a data harvest script, and the frequency and timing with which data harvesters operate. The harvest timers are controlled by additional "start_harvester" and "stop_harvester" commands.

The standard data harvester file is configured to generate publication-ready web pages with links to data and customized plots that let users do preliminary data exploration. However, in order to actually publish the data, information managers must edit two sets of files, all available in the demo directory of any standard GCE Toolbox download. There are two Matlab script files and six standard xml and css stylesheets:

- The Matlab .m files are harvest_info.m and harvest_plot.m:
  - The harvest info file contains parameters for the web file paths and links that should be established as part of the data harvest process.
  - The harvest plot file contains information on the content and type of plots or graphs that should be generated for publication.
- The look and feel of websites generated during a data harvest are managed with the following XML and stylesheets:
  - harvest_details.xsl
  - harvest_index.xsl
  - harvest_plots.xsl
  - harvest_webpage.xsl
  - harvest_details.css
  - harvest_webpage.css

By default, the harvest_info, harvest_plot, and stylesheet templates are configured to work with the demonstration data included in the Demo folder within the Toolbox. However, the harvest_plot and harvest_info files are based on case statements that can easily be copied within the file to expand the number of datasets the info and plot files are generating. These two files manage all configured harvests. Editing the stylesheets should be standard fare for any LTER information manager.

Summary and Conclusions

As we have worked to upgrade the CWT LTER sensor network, we have found that the solutions provided by both LoggerNet and the GCE Toolbox have accelerated the pace at which we can make data available. To date, we have reliable streaming data for eight sites, all of which are available for download at http://coweeta.uga.edu/streaming. As we move forward, we will not only be adding more stations, but will also be increasing the complexity and sophistication of post-processing at each site to ensure maximum data quality. We believe all of these goals can be accomplished with the framework we have outlined.

In addition, one of the Coweeta LTER PIs recently observed that, in adopting the tools we are now using, we have essentially created a "wall plug" that any CWT LTER investigator can use to their advantage when integrating sensor data into their research. Given the amount of effort that would be involved to do the same thing without these tools, we concur with this assessment and are pleased with the increases we are seeing in our capacity to scale up in order to meet growing demand.

Integrity Checks For Large Scale Databases

edit

Kali Abel

Editor's note: Kali Abel runs a small data management company based in Alaksa named Turnagain Consulting. As a fellow data afficianado we have invited Kali to submit content to this issue of Databits, both as an introduction as well as to provide the perspective of outside commentary.

I recently worked on a project that involved a large influx of data coming in from a 620,000 square mile site, derived from 900 different research work plans, 1500 researchers, and a preposterous number of samples and resulting data. Those data flooded in beginning 4 years ago and continue to come in steadily; all the while additional notes, instrument records, photos, coordinates, and corrections continue to be added. Although one could spend hours thinking about the technological processes that go into entering, storing, distributing, and describing such a large data set, perhaps the most interesting aspect of the project is the challenge of preserving data integrity when those data come in such large quantities.

For this particular project we designed and implemented various analyses that we termed "integrity checks". Because it was unrealistic to analyze each sample for accuracy due to the enormous size of the data set, it became important to find a way to screen for errors in a more efficient manner. Nearly all samples coming in were linked with a sample time and a coordinate location. Even when sample types varied, descriptive information of each sample was consistent. As can be common with field data, errors often came from transcription between field collection forms and data entry into a database form. This can mean something as simple as a latitude having two numbers reversed in the decimal degrees, or a time being recorded as AM instead of PM. To create robust data sets, screening for these errors and proposing corrections to the researchers was essential.

One such integrity check focused on reasonable velocities. For samples collected from a boat, a reasonable travel velocity was considered to be 20 miles per hour or less (given the information of the kinds of boats being employed for sampling). Using simple coding and a little bit of math, we were able to determine whether the time and distance between successive samples during a trip required a reasonable velocity. To do this we looked at the time between successive samples and the distance between coordinates using the Haversine formula (which calculates distances between two points on a sphere). All instances where velocities exceeded 20 mph between successive sample locations were flagged for further review. As a result, 498 samples were flagged out of a set run of nearly 100,000.

Additional integrity checks included an "instantaneous displacement" query which flagged all samples that had the samplers in two locations more than half of a mile apart at the same moment in time, and an average displacement query which flagged all samples that were over twice the standard deviation of average distance between all locations in a trip.

These integrity checks flagged numerous samples that could have otherwise been easily overlooked, potentially creating a domino effect of compounding errors as the data is searched and used by other researchers in their work. Integrity checks such as these allow large data sets to be analyzed for consistency more efficiently and more systematically than can often be done with large data sets. The design is often simple but the outcome allows for more interaction with research groups and a more robust and reliable data set.

Data Package Inventory Tracking: Slicing and Dicing with SQL

edit

M. Gastil-Buhl (MCR), Margaret O'Brien (SBC)

Like all LTER sites, MCR and SBC have a need for an inventory of datasets. We use inventories to:

- Build tables for annual reports
- Highlight time-series datasets that are ready to update
- Keep site leadership abreast of IM-team activities
- Plan activities for an assistant

This used to be a paper folder of Excel printouts, distilled from colorful grids maintained on a whiteboard with markers. To share status with our team we maintained HTML pages manually. That system worked all right for one person actively working on only a few datasets from a small inventory, but soon it became clear this really needed to go into a database. Coincidentally, at MCR we were loading metadata into Metabase at the same time we were preparing our inventory of datasets for PASTA, and both tasks involved keeping track of many aspects of each dataset. Patterns emerged which suggested a database design which could replace our manually-built web pages and report-tables with script-generated content. So in the summer of 2012 we loaded our content into tables and then gradually let a design evolve to meet real practical needs for both sites.

The database models from GCE LTER, which we loosely refer to as "Metabase" contain a place for just about any item of information an Information Management System could imagine. So we were surprised to discover one thing was missing: a system to track the maintenance status of all datasets in a site's inventory. We seized this rare opportunity to contribute something back to GCE, and added an additional schema within our Metabase to house the inventory tables and views. A few pre-defined queries made common summaries accessible over the web (Figure 1).

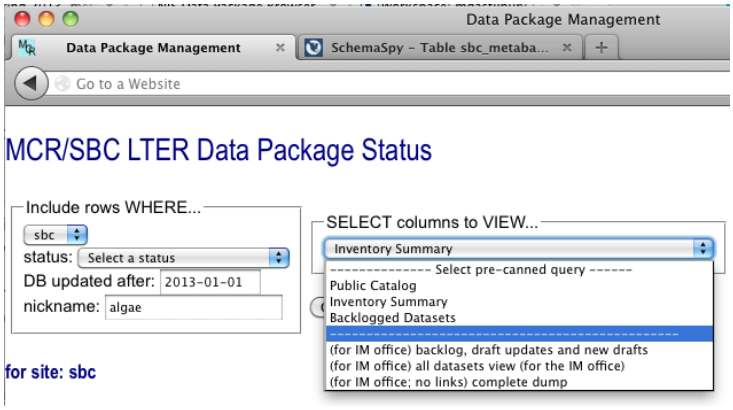


Figure 1. Web page query builder.

Now that we have database-driven inventories of our dataset packages' status, we find we are using it nearly every day. The process of designing and populating that database has brought awareness of the variety of data we handle, their specific needs for maintenance, and the details required to keep track of the data package life cycle (Figure 2). And for the first time, we have an efficient way to sort our datasets by all those categories created by the network, our site, or ourselves, such as "signature" and "core".

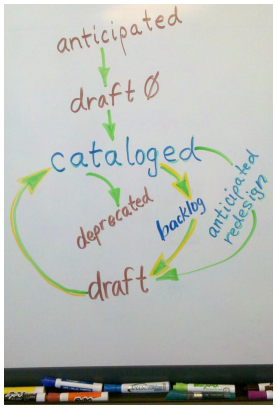


Figure 2. Package status cycle at SBC and MCR

While designing the package management database, we needed to accommodate both SBC's and MCR's management styles. We also had heard interest from some other sites in possibly using this database, so we knew we needed to keep the model general enough for wider adoption (Figure 3). We defined our stages in creating and maintaining datasets as a mutually exclusive exhaustive set, granular enough to be meaningful for different audiences and purposes. The data tables are not site-specific. The SQL views allow for some customization, and the web script which uses those views is further customized, e.g., by making hyperlinks specific to a site. The SQL views provide a crucial layer of abstraction between applications and the database tables.

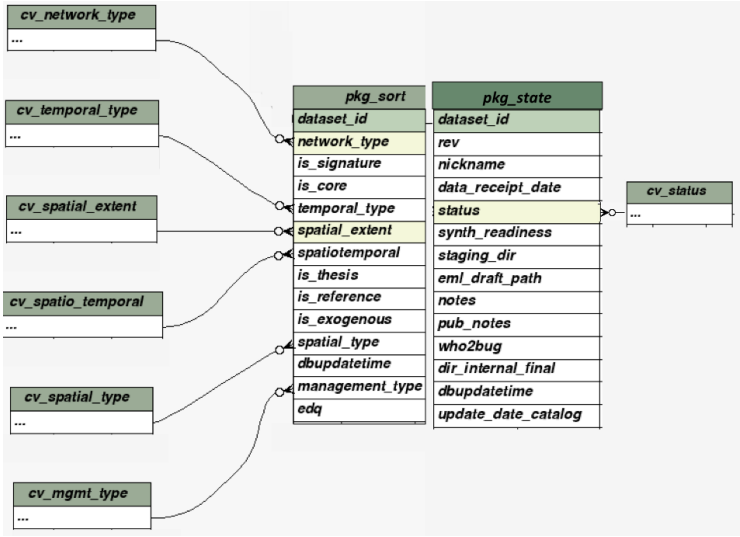


Figure 3. Entity-Relationship Diagram of data package management database schema

We have already used the database-driven queries to generate inventories for our Information Management Plans, and we expect it to streamline production of tables for reports and proposals. In addition to replacing those manual tasks, we also offer views tailored for specific audiences where the user selects from a drop-down menu. For example, a site scientist or reviewer may be interested in knowing which datasets are "student thesis data". Our site leadership asks what datasets are "backlogged" or how many have been updated in the network catalog since a given date (Figures 4, 5).

draft link	public link	status	latency	nickname	data received	catalog last updated	temporal type	who to bug	network type	title
--	knb-lter-mcr.10	backlog	1 day	Water Colum	2013-05-01	2013-01-20	ongoing timeseries	aalldredge	I	MCR LTER: Coral Reef: Water Column: Nearshore Water Profiles, CTD, Primary Production, and Chemistry
--	knb-lter-mcr.4001	backlog	1 year 1 mon 5 days	Coral Recruitment	2012-03-28	2012-02-23	ongoing timeseries	pedmunds	II	MCR LTER: Coral Reef: Coral Community Dynamics: Coral Recruitment

Figure 4. Inventory query result showing datasets with status "backlog".

public link	dataset_id	network type	temporal type	is signature	is core	is thesis	is reference	is exogenous	title	data last received	catalog updated	status
metadata and data	knb-lter-mcr.1	I	ongoing timeseries	t	t	f	f	f	MCR LTER: Coral Reef: Community Dynamics: Abundance and Species Richness of Fishes Associated with the Coral Porites rus	2012-04-16	2013-01-20	redesign anticipated
metadata and data	knb-lter-mcr.10	I	ongoing timeseries	t	t	f	f	f	MCR LTER: Coral Reef: Water Column: Nearshore Water Profiles, CTD, Primary Production, and Chemistry	2012-03-13	2013-01-20	backlog
metadata and data	knb-lter-mcr.1034	I	ongoing timeseries	t	t	f	f	f	MCR LTER: Coral Reef: Water Column: Nutrients	2012-03-13	2013-01-20	revision pending
metadata and data	knb-lter-mcr.1035	I	ongoing timeseries	t	t	f	f	f	MCR LTER: Coral Reef: Benthic Water Temperature	2012-02-28	2013-02-14	cataloged
metadata and data	knb-lter-mcr.1036	I	ongoing timeseries	t	t	f	f	f	MCR LTER: Coral Reef: Bathymetry Grid for North Shore	2011-11-23	2013-01-20	cataloged
metadata and data	knb-lter-mcr.1037	I	ongoing timeseries	t	t	f	f	f	MCR LTER: Coral Reef: Water Column: Offshore Ocean Acidification: Water Profiles, CTD, and Chemistry	2012-03-14	2013-01-20	cataloged
metadata and data	knb-lter-mcr.1038	I	ongoing timeseries	t	t	f	f	f	MCR LTER: Coral Reef: Long-term Community Dynamics: Backreef (Lagoon) Corals Annual Survey	2012-02-11	2013-01-20	cataloged

Figure 5. Inventory query result showing categories (only top few shown here).

Metabase Adoption by SBC and MCR

edit

Margaret O'Brien (SBC), M. Gastil-Buhl (MCR)

Metabase is a comprehensive relational database designed in Microsoft SQL Server, which has been in production at GCE LTER for many years. It already has all the features needed for housing and exporting EML metadata, and work is ongoing to integrate EML with other tools at GCE, most notably the GCE Matlab Toolbox (Sheldon, et al. 2012). Two LTER sites co-located at UC Santa Barbara, Santa Barbara Coastal and Moorea Coral Reef (SBC and MCR, respectively) required a robust relational database to underlie many features of their information management systems (IMS), including web pages for people, research projects and dataset views, and chose Metabase for this (Gastil-Buhl, 2010). We began our adoption process by porting Metabase to the open source relational database management system PostgreSQL in 2011, and the first exports have been from the research project tables as lter-project XML (O'Brien, 2011). Our next milestone is export of EML datasets. This article provides an overview of our recent progress.

Populating Metabase

Each site has its own Metabase instance in PostgreSQL, and the GCE core model was installed into three PostgreSQL schemas: 'metabase2', 'research_project' and 'resources', which are analogous to two of the SQL Server databases in GCE Metabase (Metabase and Submissions, Sheldon, 2011). The views used by GCE were also ported, and serve as an abstraction layer between the model and export scripts. In many cases we will use Metabase views as designed. However, in some cases we will create additional views, especially when use of PostgreSQL XML type is indicated. We created a separate schema for these views, called 'trees'. The research_project schema was populated first, and we are now working mainly on datasets, in the metabase2 schema.

Datasets use many more tables than do research projects. Both MCR and SBC have rich, high quality EML metadata, and routinely use most dataset module elements. Our strategy was to use this already-structured metadata to populate Metabase. Across multiple static EML documents, content is not constrained, allowing some fields to be irregular. An opportunity to regularize this content arises during upload to Metabase. Within a relational database, constraint and regularization of content is built in, and in a normalized database, the parent tables must be populated first.

As of this writing, all dataset-level content for both sites is populated, and we are progressing with the entity-level metadata. A general pattern for table-population has emerged:

1. Ascertain the population-order for a group of tables with a schema analyzer such as SchemaSpy.
2. Using XSLT, extract the subset of EML content needed for the table(s) from the entire corpus of EML documents to flat text files.
3. Upload content to a generic 'scratch' table and regularize content as appropriate.
4. Add primary and foreign keys to the scratch table to ensure constraints are met.
5. Use SQL to copy content from the scratch table to the respective metabase tables.

Our scripts for extracting database table content from EML are reusable to some extent. With continued use, we have been able to factor some components to make them more generic.

Collaborative Cross-site Design

We used existing GCE Metabase content as a guide, and as much as possible populated SBC/MCR tables in the same way. In some cases, however, the model did not quite fit our local concept of a dataset feature. This is understandable; Metabase has been in production since 2001 with multiple uses at GCE, and evolved to suit the needs of that site. We described each issue we encountered, and have categorized our approach to solving them:

1. Solving the incompatibility is essential to continuing, and our solution to the issue IS NOT backward compatible: we wrote up a proposed change and contacted Metabase's author. These issues were most important to discuss with GCE before proceeding.
2. Solving the incompatibility is essential to continuing, and our solution to the issue IS backward compatible: we proceeded with the best solution for SBC and MCR, wrote up our solution and planned to share with GCE at a future date. Possibly, our solution would be more broadly applicable, but we thought it more considerate and efficient to work out potential cross-site solutions in a more formal manner.
3. Non-essential changes, whether backward-compatible or not: we made notes, planned to bring it up later, and moved on! These instances are just alternative ways of modeling the same information.

Conclusion

NSF leadership has encouraged us to leverage existing tools within the network, and the GCE suite of tools has a long track record of continual upgrades to meet increasing expectations, making it an ideal candidate. In our work, we have invested extra time to carefully design code enhancements that can be merged back into the shared model. In developing additional export scripts, a major focus is also on reusability by other sites. Additionally, we are taking advantage of the SQL/XML features which are now available in PostgreSQL 9, but which were not available in SQL Server in 2001 (when Metabase was originally written). GCE has expressed interest in porting these enhancements back into SQL Server, and we hope they meet the high standards exhibited by that site.

References

Gastil-Buhl, M., et al., 2010, Metadata database models and EML creation, poster at IMC annual meeting. http://im.lternet.edu/node/694

O'Brien, M. 2011. The Santa Barbara Coastal implementation of projectDB using Metabase. http://databits.lternet.edu/fall-2011/santa-barbara-coastal-sbc-lters-implementation-projectdb-using-metabase

Sheldon, W.2011, GCE Metabasew Database Schema, https://gce-lter.marsci.uga.edu/public/app/resources.asp?type=document&category=informatics&theme=Database%20schemas

The New Drupal Ecological Information Management System

edit

Inigo San Gil (MCM)

The overarching goal for Drupal Ecological Information Management System (DEIMS) is to make information management simpler. In other Databits articles the DEIMS working group explained why we choose Drupal, the same platform that the White House and US House of Representatives currently use to manage their information streams: it is free, widely adopted, forward-looking, and open for easy customization. In a nutshell, it is a platform candidate for "IM-in-a-box", a Swiss Army knife for information managers. Right now the most active members in the DEIMS working group in the US-LTER network range from pole to pole, including the nothern-most Arctic LTER, the torrid Jornada, tropical Luquillo Experimental Forest, the tranquil North Temperate Lakes, the south polar deserts near McMurdo, the Plum Island Ecosystem north of Boston, the north end of the Chiuahuan deserts at the Sevilleta LTER, and the headquarters in Albuquerque, NM. The DEIMS project also benefits from contributions from our international partners at ILTER and the University of Michigan Biological Station.

In this article, I would like to brief you about what to expect in the upcoming DEIMS release, which should hit the shelves Summer 2013. DEIMS has been progressing though various stages of development through our working group for about four years, and we are now in the midst of work on the next release of this system that we use to manage data and information to produce the best site-based knowledge we can foster. Over the best part of last year, the DEIMS working group labored at securing a contract with one of the top development shops in the Drupal community. That process was not pretty; we had to coordinate a multi-institutional effort, create a public RFP competition, engage lawyers from four universities, and conduct diligent reviews of the proposals submitted. In the end we had to pick one proposal, which was not easy as we had five very strong candidates. After counting votes and validating reviews, we are proud to announced that Palantir.net came in at the top. We secured a contract with Palantir after another grueling round of paperwork and deciphering legalese. NSF supplment contributions from NTL, SEV, and LUQ, with support from the Network Office, made this contract possible. Kudos to all supporters. This is the first time DEIMS has had some non-negligible level of network financial support for development, certainly a sign of progress.

But if Drupal is free and open, why hire a contractor? Although the DEIMS working group members have gained in-depth Drupal expertise, we thought that it would be good to get a specialized, professional review of the early DEIMS work adopted by the SEV, LUQ, NTL, UMichigan Bio Station, and Europe-LTER. Several Drupal professionals gave the current DEIMS product thumbs up. While we were at it, we also decided to ask the Drupal professionals "what Information Management Package would they develop if they had to do this again, with a given set of requisites and requirements". Our team took the challenge, and here we added their ideas and suggested improvements --read a brief list of what you can expect.

Network Compliance

- DOI retrieval, PASTA compliance mark
- Integration of the LTER Unit dictionary services
- Controlled Vocabularies from diverse authorities

Easy Searching

- Faceted Search
- Contextualized Data Search
- SEO engineering

Services

- Offer a service for EML and PASTA compliant metadata
- Other export services, including RSS and flat text

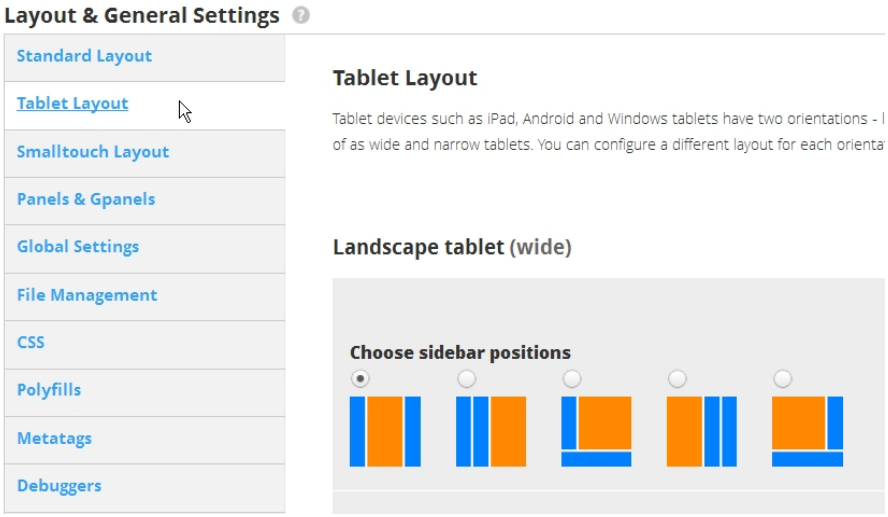
User Interface

- Browser-based information management
- Dynamic database-driven forms
- Mobile-friendly website
- Role based webflows for data and metadata chain of custody

Detailing all these points is beyond the scope of this article, but we will touch on three of the above.

Mobile-friendly. We selected the pixture-reloaded theme, which provides an adaptive interface out of the box. All the default configuration settings fit your web pages to any display used to render them, whether small smartphones, tablets, or big screens.

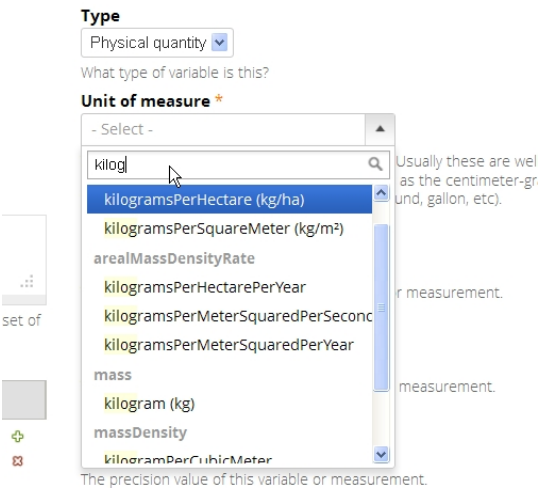
The figure below is a partial screenshot of the look and feel of configuration options; what you see is the selection for the "Table Layout"



Note that we do not use device detection (a futile exercise in my opinion). Instead we use display detection for adaptation strategies.

Like in the DEIMS product we currently use, we are bundling a slider (a carrousel, slide show: cycling slides w/ news noteworthy stories), the difference, is that this one is adaptive, and resizes itself nicely to the contour conditions (screen or browse size).

Network Services: DEIMS will leverage DOI retrieval, PASTA compliance status, EML service, controlled vocabularies and the LTER Unit Vocabulary. In addition to using the Unit Services to make value added descriptions of units using the STMML format, in this new DEIMS version we provide a form connected to the Unit Service that suggests the appropriate units, while preserving the ability to manually enter a new unit. In addition, DEIMS will cache the unit dictionary results, in case the service or the connection is interrupted. Below is a relevant screenshot, which shows the form field for units with the Unit Dictionary lookup.



Faceted Search. The ability to narrow down result search sets is paramount for specialized searches. Google-like searches that retrieve relevant results that often miss the target of the desired search, e.g, a specialized ecological keyword, such as "primary productivity" may retrieve results from the site bibliography, the photo gallery, the datasets, the research projects, even personal profiles. DEIMS will offer the ability to refine the search results with a series of contextual filters. In addition, the default native database search engine can be swapped with the powerful Apache Solr search with an easy pull-down menu from the configuration pages.

Chances are you have used the Faceted Search functionality on many e-commerce sites, including Google shopping, Amazon, and about any serious retailer. The Facets process enables you to restrict or refine your results to specific brands, price ranges, or even customer feedback or rating. In DEIMS, we can restrict the searches by date of publication, date of study, location, type of information (dataset, publication, etc.), or even the keywords assigned to tag the content.

Where and when can I try the new DEIMS? The new release of DEIMS will be available for download at drupal.org as well as at GitHub. Come see a demo at the next Watercooler or next meeting. If you cannot attend, you only need to request any DEIMS team member for a demo, or access to any site. We are plaaning new exciting trainings and features to further this work. For example, we are working hard to bridge ecology with molecular genomics (microbial and fungi overwhelming biodiversity). We set our aims on killer data visualization apps.

Commentary

My Experiences as a Participant in the Sensor Training Workshop

edit

Fox Peterson (AND)

The sensor software tools training workshop held the week of April 22, 2013, at the LTER Network Office in Albuquerque was outstanding. I wanted to write up a brief overview of my experiences as a participant in hopes of summarizing the experience and exciting others about participating in data management workshops in the future. A recurring theme throughout the workshop was that of "middleware"; that is, a tool that bridges raw data input to database storage in such a way as to provide standardized quality control and reduce the need for redundancy in writing programs or repeating processes to consolidate data. Thus, I'll focus on that here, and how I think I will use it in my work with the Carbon-budget group, Biomass group, Biomicrometeorology lab, and DIRT (detritus input and removal treatment) network at H.J. Andrews (AND).

On the first day of the workshop, we learned about the CUAHSI Observations Data Model (ODM) and associated data import and visualization tools in the Hydrologic Information System. Basically, the ODM tools are capable of taking raw data, assigning relevant metadata, performing user specified quality checks, and organizing that combination of metadata and data in a database (we used Microsoft SQL Server). What was fantastic about this tool to me was that once the metadata and data were joined, this process never needed to be repeated unless a new sensor was added. Although I don't generally work with hydrologic data, I tested the tool on data from two sonic anemometers (one in Antarctica and one at the H.J. Andrews) and was able to reuse the metadata values shared by both anemometers while specifying those values specific to site. This was very convenient! Another great feature of this tool for those who do work with hydrologic data is that the CUAHSI HydroDesktop can import directly from several key stream gage websites, such as USGS sites, and put the extracted data into a database. I have searched USGS stream data before and been annoyed at having to click on the links of multiple gages, download the data by hand, and make it into a format I can use; using the HydroDesktop would be a huge time saver for scientists in fields like urban aquatics.

On the second day, we were fortunate to have Matt Miller of Data Turbine come and speak. Admittedly, I am not sure my skills in "networks" were quite to the point that I fully benefited from his talk, but I was able to get the basic gist of it and play with the Data Turbine tools. Data Turbine essentially acts as a harvester between users and streaming data, and moreover they are sharable and controllable through the Ring Buffer Network Bus (RBNB), which offers both added connectivity and added security for data assimilation. Matt showed us how to set up a local server and download sample data from the North Temperate Lakes LTER, and then plot and quality check this data in near real-time. He was very attentive to the importance in some fields of not only having extremely fast data but having access to this data as soon as possible, and to how memory and power constraints come into play with this type of operation. I feel that I will be able to use what I learned from Matt to think about our meteorological data for the biomicrometeorlogy group and C budgeting on AND Watershed 1. The second afternoon's lecture was my favorite because it was about my favorite tool, MATLAB! Wade Sheldon, LTER's reigning MATLAB expert, introduced us to the GCE Data Toolbox and helped us develop data templates for our own loggers, as well as play with some existing data he had from a Campbell data logger. I could talk for quite some time about this tool, but mostly I am eager to use it for all of my own data imports, develop templates, and share these templates with Wade. I would love to become a part of the development of the GCE Data Toolbox. Additionally, some of the functions which run with the GCE Data Toolbox are extremely helpful; Wade has modified some functions such as the Matlab "whos" to do quality checks or locate errors in imports; this will greatly expedite data pre-processing!

On the third day, we began with a talk on Kepler. I will honestly say this is probably the tool that I will end up using the most of all. One great thing about Kepler is that it integrates rather seamlessly with R and can be used to generate fairly open ended R-based workflows. I kept thinking about most of my data with Biomass, DIRT, and C-budget. This data is not traditional "sensor network" data, but it is big, and messy, data; if we were to view humans collecting observations as a sort of sensor that gather information irregularly, then this observational data fits within the "sensor network" framework. Many humans gathering data are not comfortable with the use of tools like SQL Server or MATLAB, and prefer spreadsheets. However, if they were given the opportunity to simply make a flat file and turn on a pre-defined Kepler workflow to import this spreadsheet into a database like SQL Server and produce meaningful output (like graphs or analyses in R), I believe they would be accepting of this. This would be fantastic for DIRT - for example we could standardize metadata and data storage across scientists and sites by running the same workflows. That Kepler and R are both open source is also a huge plus, as MATLAB and SQL Server can be too "spendy" for some! The afternoon on the third day was devoted to R, and that was good practice for me, because I prefer MATLAB and always need time to brush up on my R in order to stay fluent.

On the final day, we had an excellent presentation on the National Ecological Observatory Network (NEON) from Jeff Taylor. All I can say is, *wow*, if I was fired up about sensor networks at the site level, I'm even more fired up about site networks at the synoptic level. NEON is a huge dream of standardization, homogenization, and synthesis, with the ability to produce some very powerful results that may be able to affect non-science consumers. As a member of the C-budget group and originally a forester, I believe that the arguments something like NEON could put forth for important topics such as climate change and forest restoration would absolutely make a huge impact for many people. I had the good fortune to talk to Jeff afterwards and share emails, and his soil scientists and I are going to speak about the standard ways of sampling soil stores and effluxes in the footprint of eddy covariance towers. This will be super because I will get to share with them the challenges of working in complex terrain, and they will be able to help me establish a framework of what is the "norm" in non-complex terrain. I can't wait to be more involved with information management!

I want to give a HUGE thanks to Don, Corinna, and the many trainers. I am also fortunate to have met many new "nerdy friends" who I know I will be drawing on in the future. This is always an added benefit of a well-run workshop; that one comes away not only with new tools, but with a big, extensible, friendly "support team". Learning about different types of middleware to automate metadata assignment and quality control makes me more and more comfortable with larger and faster data, and I am eager to see how data management will evolve in the next few years with the availability of these fantastic tools. THANKS EVERYONE!

I Was Told There Would Be Cake

edit

Kali Abel

Editor's note: Kali Abel runs a small data management company based in Alaksa named Turnagain Consulting. As a fellow data afficianado we have invited Kali to submit content to this issue of Databits, both as an introduction as well as to provide the perspective of outside commentary.

That, in a nutshell, is quite possibly how I become a data manager.

I have been a research scientist for about 7 years, finding myself on projects all over the world, one leading to the next, one idea or question transforming into another. So in some ways I guess it made sense that entering the data management world happened in a similar fashion. I got asked to join a project that was vastly different than any other I had been on, requiring me to be a data manager overseeing the data from thousands of studies spread across hundreds of miles, from hundreds of different scientists, researchers, and volunteers. As a researcher I had always shepherded projects all the way through their life cycles, from data collection to data management to reporting and accessibility for end users. My passion is in finding ways to convey the data to a varied audience, and making my research more accessible to colleagues and general public. Therefore I cannot truly justify any surprise when I came for the cake and stayed for the party.

I cannot ever overstate how important communication is in science, and so much of that communication is done through how we organize, display, understand, and convey the data generated through research. Yosemite National Park's recent research on climate impacts on high alpine glaciers is the perfect case study in communicating data – and not just any data. Field research provided the data you might expect: glacier measurements, sensor data, meteorological station data streams, mass balance calculations, tree cores, and stream flow measurements. But that wasn't even the beginning. Prints going back to the mid-1800s, articles from Harper's Magazine written by John Muir, maps drawn by IC Russell, USGS markers from World War II, 70 years of Parks Service trip reports, photos and sketches and stories and notes and drawings in the margins. Not to mention the stories told by those familiar with the glaciers and the questions asked by those who were intimately familiar with the changing landscape. With the right understanding and management, all of these pieces could somehow fit into a more complete image, adding incredible depth to the research. Such a huge amount of data, in so many forms, demanded attention and a coherent form of communication. And that was only step one.

Step two was to convey such a story to scientists, researchers, the public, park managers, and volunteers. This required making the data useable and approachable for a vast audience without diluting or overcomplicating the message. And because it's infeasible to address everyone at once, we must find ways for each user to get what they need from the data through an interface that doesn't require any additional support. The platforms for communicating the data suddenly become as numerous as the packages the data originally came in.

I'm passionate about communicating data to anyone, be it researchers, scientists, other data managers, students, policy makers, the public, or guests at a dinner party (sorry, Mom). It affords me a way to communicate the science that pulled me in originally and challenges me to be constantly transforming and innovative in the quest to build the link between scientists and end users. As the field of data management grows, I am thrilled to be growing with it and finding ways to communicate science with any audience that wants to learn.

Please pass the cake.

News Bits

GCE and CWT Host Successful Workshop to Demonstrate, Improve, and Promote the Adoption of the GCE Data Toolbox for Matlab

edit

John Chamblee (CWT), Wade Sheldon (GCE), Richard Cary (CWT)

As the volume and diversity of ecological data grows, scientific discovery demands ecological scientists and anthropologists develop common tools to solve common problems so that data, as well as published literature, can be used to frame and envision next-generation research. From November 27-30, 2012, the Coweeta (CWT) and Georgia Coastal Ecosystems (GCE) information managers pursued this goal by leading a workshop on the GCE Data Toolbox for MATLAB. At this workshop, information and data managers from 11 universities and federal agencies were provided a potentially critical step in meeting the need for a common set of tools. The workshop was organized so that attendees were offered time for hands-on instruction that not only provided an introductory framework, but also a considerable amount of unstructured time in which information managers could interact with the software and its developer using their own data to solve their own problems.

Workshop Participants, Back Row (L-R); Wade Sheldon (GCE), Vincent Moriarty (MCR), Sven Bohm (KBS), Dominik Schneider (NWT), Aaron Stephenson (NTL), Kristin Vanderbilt (SEV), Chris Gotschalk (SBC), Margaret O'Brien (SBC), John Porter (VCR), Richard Cary (CWT). Back Row (L-R): Adam Sapp (GCE), Marty Martin (HBR), M. "Gastil" Gastil-Buhl (MCR), Hope Humphries (NWT), Inigo San Gil (MCR), Stephanie Laseter (USDA FS), Adam Kennedy (AND), John Chamblee (CWT)

The workshop was co-led by Wade Sheldon (IM at GCE), John Chamblee (IM at CWT), and Richard Cary (Assistant IM at CWT). Sheldon provided the vast majority of the instructional support, with Chamblee providing logistical support, instructional assistance, and an instructional workspace at the CWT IM Office, which is housed in the UGA anthropology department's Sustainable Human Ecosystems Laboratory (SHEL). Cary and Sheldon produced a step-by-step manual for many Data Toolbox core functions.

LTER information management staff from AND, HBR, MCM, MCR, NTL, NWT, SBC, SEV, and VCR attended the meeting, as did additional IM staff from GCE and the Coweeta Hydrologic Laboratory. The workshop consisted of about 2 half-days of formal presentation, followed by half-days of one-on-one work in which attendees used tutorials and sample data sets (on the afternoon) and their own data (on the second afternoon) to explore and test the Data Toolbox as a means for meeting their needs. The final morning of the workshop included an intensive discussion about potential drawbacks of the toolbox and on gathering data for adding functionality and making improvements.

This final discussion resulted in the collection of several key suggestions, some of which have already been implemented. These include a listserv for GCE Data Toolbox users that, since its inauguration, has proven an active and valuable source of community input. A post-workshop survey suggested that most of the attendees plan to integrate the toolbox into some aspect of their site information management activity and that, in areas where the toolbox will not prove to be useful, this is often the case because good solutions are already in place. The survey also demonstrated the success of a mixed-format approach to instructional presentation and highlighted the importance of presenting tools that were mature enough to allow users to test them with their own data.

As Sheldon continues making technical improvements based on feedback from the workshop and the LTER community, the presenters plan to expand the print-based tutorial materials to accompany these changes and to release short tutorial videos.

All of the instructional materials from the workshop are available on the GCE Data Toolbox Documentation Page (https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox/wiki/Documentation). Users who wish to join the GCE Data Toolbox listserv can email any of this article's authors. The workshop was primarily funded through a contract received from the LTER Network Office using funds designated from American Reinvestment and Recovery Act of 2009. Mathworks, the maker of MATLAB, also supported the workshop and its goals by providing consultation and short-term licensing support. A final report on the workshop and associated software development efforts will be publicly available some time shortly after July 1, 2013.

Sensor Networks Training Conducted at LNO

edit

Don Henshaw (AND), Corinna Gries (NTL)

The "Software tools and strategies for managing sensor networks" training workshop was conducted at the Long-Term Ecological Research (LTER) Network Office (LNO) on April 23-26, 2013. Funding was provided by the LNO with cost-sharing from Tony Fountain's DataTurbine group at the University of California - San Diego (UCSD). Twenty-three participants and eight trainers were on hand representing a very diverse mix of researchers, students, technology experts, and information managers from over 20 domestic and international institutions and including affiliations with 10 LTER sites.

This training workshop was intended to introduce participants to a variety of software tools that are useful in the transport, quality control (Q/C), management, analysis, and storage of data streams from sensor networks. Hands-on exercises were provided for each tool and several students were able to successfully employ their own data logger (.dat) files. Jeff Horsburgh (CUAHSI) introduced the CUAHSI Hydrology Information System including the HydroServer and HydroDesktop. Participants learned how to prepare and map their data into the Observation Data Model (ODM) for storage and access through CUAHSI tools. Sameer Tilak and Matt Miller presented DataTurbine (DT) and its installation and trainees were able to pull data from an active North Temperate Lakes (NTL) DT server, graph the data for quality checking and explore with the Real-Time Data Viewer (RDV). Wade Sheldon presented the GCE Data Toolbox and highlighted many features including Q/C, metadata management, and visual graphics. Wade also illustrated the integration of software tools by demonstrating the import of both Campbell data logger and data from a DT server as well as the export of this data into the CUAHSI ODM. Corinna Gries presented an overview of Kepler and hands-on exercises to build analytical components for importing data from a DataTurbine server into Kepler. John Porter followed via videoconference with an introduction to R and examples of analyzing sensor data using R within Kepler. Demonstrations showing the integration or "chaining" of these software tools was intended to provide continuity over the 3.5 days of training. The workshop agenda, presentations and additional materials are posted (http://im.lternet.edu/im_training/sensors_2013).

Consideration of strategies for managing sensor networks was a secondary theme to the workshop. Jeff Taylor (NEON) presented an overview of NEON with a focus on Q/C and publication of NEON data. Don Henshaw presented some of the preliminary results from an earlier workshop, "Sensor management best practices", which is part of a community effort to summarize problem statements, best practices and present use cases on a variety of sensor network issues. This work will be posted on a wiki hosted by the Federation of Earth Science Information Partners (ESIP) (http://wiki.esipfed.org/index.php/EnviroSensing_Cluster).

Overall, the initial feedback for this workshop was very positive. Favorable comments were expressed for each of the presented tools, with perhaps the GCE Data Toolbox for MATLAB receiving broadest acclaim. Please see the editorial in this DataBits issue written by participant Fox Peterson on her workshop experience.

This is the second cost-shared sensor training workshop co-organized by this article's authors following up on a similar workshop funded by the LNO, NCEAS and DataONE in May 2012. Once again there were far more applicants than spaces available for this training. LTER and other ecological sites are actively developing sensor networks and are looking to share or adopt technological solutions. This training session, along with the concurrent effort to describe best practices for sensor network management, represent great examples of the LTER reaching out and working with the broader ecological community to improve efficiency and standardize approaches in managing sensor networks.

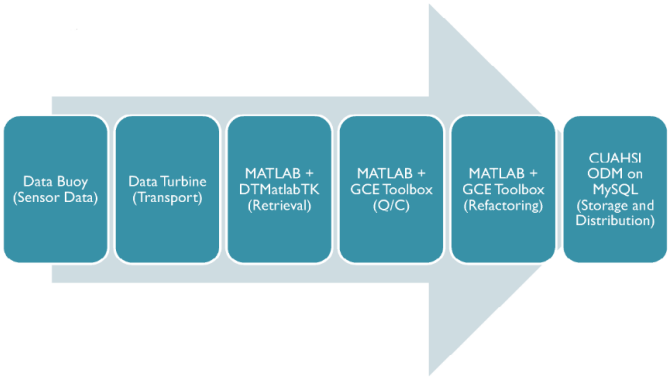
Integrating Open Source Data Turbine with the GCE Data Toolbox for MATLAB

edit

Corinna Gries (NTL), Wade Sheldon (CGE), Tony Fountain, Chad Sebranek (NTL), Matt Miller, and Sameer Tilak

North Temperate Lakes LTER's streaming sensor data are being used as one of three "science experiments" in a NSF *Software Infrastructure for Sustained Innovation (SI^2)* project led by Tony Fountain (CalIT2, UCSD). A major focus of this collaborative project is software integration in complex science environments. This involves strategic software development, systems integration, and testing through demonstration projects (i.e., science experiments). Major requirements for the software developed by this project include performance, usability, interoperability, and cyber-security. In addition to NTL LTER, these software products will be integrated into production research infrastructures at Purdue University, the University of Connecticut, and the Monterey Bay Aquarium Research Institute to answer important science questions, including: *(1) What is the impact of uncertainty in the design of civil infrastructure? (2) How sensitive are ocean systems to pH changes? (3) What is the variability of lake metabolic parameters such as gross primary productivity and respiration?*

One goal of this collaboration is to make integrating the Open Source Data Turbine (OSDT) streaming data middleware with other environmental community software tools more robust and accessible to information managers. In the first project phase, the existing OSDT - MATLAB interface was improved by developing a toolkit (DTMatlabTK) of easy-to-use MATLAB functions for interacting with Data Turbine servers. Building on these improvements, code was developed to directly access data in OSDT using the GCE Data Toolbox for MATLAB (developed at Georgia Coastal Ecosystems LTER) to provide a robust, automated and configurable QA/QC environment for harvesting real-time sensor data. The GCE Data Toolbox was then used to transform data to be compatible with the CUAHSI Observations Data Model (ODM, see Resources section below for links), and insert processed OSDT data into an ODM database to support an end-to-end workflow from NTL data buoys to a CUAHSI Hydrologic Information Server (figure below).



The functionality developed during the first phase of the project was introduced in a recent training, titled "Software tools and strategies for managing sensor networks", held at the LTER Network Office in April 2013 (see Henshaw & Gries this issue and Resources section below).

Following this workshop we will implement the new functionality in a production level streaming sensor data management system at NTL. Our experiences will be documented in detailed user guides and tutorials. Additional developments will include creating OSDT software interfaces that are compliant with the Open Geospatial Consortium Sensor Web Enablement standards, and making improvements to the OSDT – Kepler workflow system interface.

Resources:

Home page for the training workshop "Software tools and strategies for managing sensor networks" http://im.lternet.edu/im_training/sensors_2013. Relevant presentations from that workshop: http://im.lternet.edu/node/1169 and http://im.lternet.edu/im_training/sensors_2013/gce_toolbox

NSF award information for this project: http://www.nsf.gov/awardsearch/showAward?AWD_ID=1148458&HistoricalAwards=false

Open Source Data Turbine: http://www.dataturbine.org/

GCE Data Toolbox for MATLAB: https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox

Open Geospatial Consortium, Sensor Web Enablement: http://www.opengeospatial.org/projects/groups/sensorwebdwg

Kepler Project workflow system: https://kepler-project.org/

CUAHSI ODM: http://his.cuahsi.org/documents/ODM1.1DesignSpecifications.pdf

Vista Data Vision Software Training Workshop

edit

Jason Downing (BNZ)

For those of you in the market for software solutions to better manage your complex sensor networks, one product vendor has recently developed a training curriculum to help you along the way. Along with other open-source and LTER network developed tools and methodologies with which to manage sensor networks, there are also some commercially available software options, one of which is Vista Data Vision (http://vistadatavision.com/vdv2013/). An earlier version of the VDV product was originally discussed in the Spring 2009 issue of Databits (http://databits.lternet.edu/issues/113). The developers have recently released the latest version of their product, VDV2013, and in conjunction with it have developed a training course to offer as support to their growing group of users. In early March, a call was sent out to the list of VDV users announcing that they would be offering an initial training class at their headquarters in Reykjavik, Iceland for those users who are either new to the software or would like to better exploit all the improved features this software has to offer. There was significant interest in this offering and the available slots quickly filled up. I was personally lucky enough to obtain one of those slots. We have been using VDV at BNZ since the fall of 2008 and have worked with the developers at Vista to aid in the evolution and development of their product over time. With the recent product upgrades, we felt this would be an excellent opportunity to obtain some advanced training to facilitate our transition and implementation of the newest version and its expanded features.

This initial training was offered at no cost for its participants, but that will likely not be the model as they anticipate additional offerings in the future. They did an excellent job of developing a comprehensive curriculum to cover the setup, configuration and utilization of all the various product features. One of their most significant advancements is the transition to offer more of the software functionality through a web-based interface to simplify the server administration duties and improve control and utilization from remote locations by various system users. They have also added new tool kits to provide an enhanced Google Maps functionality and some additional graphical display opportunities. They also spent time discussing what is in development for future releases and seeking additional input as to what other features are in demand.

The opportunity to train and interact with the software developers in their home environment was extremely beneficial. They have always been open and willing to assist and answer questions via email but as always, the opportunity to be in the same room and have detailed discussions and work through specific application scenarios remains priceless in comparison. In addition to the training material and hands-on exercises, they provided additional time to interact one-on-one with each of the participants to discuss their individual application issues. They provided access to and support from all of their development staff so everyone in the training was able to customize the teaching examples for each participant's particular situation or application. It was also valuable to interact with other software users and get a full appreciation of the breadth and variety of applications that the software is capable of supporting. The training was attended in large part by

people in the geotechnical field, currently the largest user group of the software, but there were also people representing the wind power and environmental science fields. As a distributor of this software, Campbell Scientific also sent a representative to gain additional training information to better support their customers (which includes many of those associated with the LTER program).

The hospitality in Iceland is well known and the folks at Vista were no exception. They proved the rule with the utmost class and extended themselves beyond expectations to provide a complete and fully rewarding experience. If others using this software have the opportunity to participate in training events such as this, I would highly recommend taking of advantage of this opportunity when it is next available from the Vista Data Vision crew. Additionally, make sure to schedule some extra time to enjoy everything else that Iceland has to offer, which is abundant.

GeoNIS Project Update

edit

Adam Skibbe (KNZ), Theresa Valentine (AND), Aaron Stephenson (NTL), Jamie Hollingsworth (BNZ), Ron Beloin

Following the 2012 All Scientists Meeting, the GIS Working Group hired a developer to help realize a working version of the GeoNIS. Using funds provided by the Network Office, the GIS Working Group hired Ron Beloin, a programmer from Oregon State University with a background in Geographic Information Systems and web services, to develop the GeoNIS workflows. In addition, funding allowed for a 2 day meeting at the LNO in January that included core GeoNIS personnel and Mr. Beloin to construct a work plan. At this meeting we worked with LNO staff to identify how the GeoNIS may best be incorporated into the NIS as a whole. Focal points included full integration with PASTA, defining the workflows needed to move data from PASTA into a working Geographic Information System, and integrated geospatial quality checks. Jamie Hollingsworth worked with LNO staff to configure and test the network server, including providing remote access for the GeoNIS team members. As a means to remain compliant with the interests of the individual LTER sites, the group decided that any dataset with restrictions would not be ingested at this time.

As described in the Sping 2012 Databits (http://databits.lternet.edu/spring-2012/geonis-adding-geospatial-capabilities-nis), the GeoNIS will provide workflow tools to extract spatial data from site level EML housed in the PASTA framework, develop web mapping services for individual LTER sites, and support cross site efforts (StreamChemDB and ClimDB/HydroDB). The project will also provide quality checks for vector and raster datasets analogous to what PASTA does for tabular datasets. The workflows being developed by Mr. Beloin use Python programs and XSL stylesheets. All processes are being developed on a server in the LTER Network Office with ESRI software (specifically, ArcGIS Server 10.1) on a PostgreSQL database. The workflows rely on EML documents to identify spatial datasets within PASTA, and to provide descriptions of the data within the mapping and image services. Linkage to the PASTA data packages and source EML are maintained within the GeoNIS.

The following workflow products were initially tested with sample EML files (not in PASTA), and have been further tested on a subset of data drawn from PASTA (KNZ and NTL scopes):

1. Discover and retrieve data inside of PASTA that has not yet been processed into the GeoNIS
2. Unpack PASTA data packages and parse the EML file
3. Perform data quality checks
4. Load vector type data into geodatabase
5. Augment metadata for vector data
6. Load raster type data into raster geodatabase
7. Augment metadata for raster data
8. Update map documents (used for creating web mapping services) and/or master mosaic datasets (used for creating image services)
9. Refresh map and image services and update layer information for queries
10. Update geodatabase with tracking and progress information
11. Compress and archive work directories; perform other cleanup tasks

Next steps include setting up scheduled polling of PASTA for new datasets to ingest, ingesting data from additional sites (moving the GeoNIS from testing to production), developing a user interface to help site personnel access the web mapping services, and developing web applications for viewing the spatial data. Provided there are no major hiccups in development, there will be an update and demonstration on GeoNIS at the 2013 LTER IMC meeting.

Once spatial data for a site is stored in the GeoNIS, researchers/students will be able to access GIS data from the map and image services (for their site and cross-site) and use them in desktop GIS software or through web mapping applications. Eventual functionality will include providing geospatial analysis products, gazetteer services, and interactive web mapping portals (like LTERMapS). The GeoNIS team will be working on helping sites use the web services, and providing additional functionality as found to be valuable. It is our goal that the GeoNIS be the geospatial backbone of the NIS, and will provide quality-checked data and services to those wishing to use these data and services into the foreseeable future.

Coming Soon: LTER Landsat Catalog

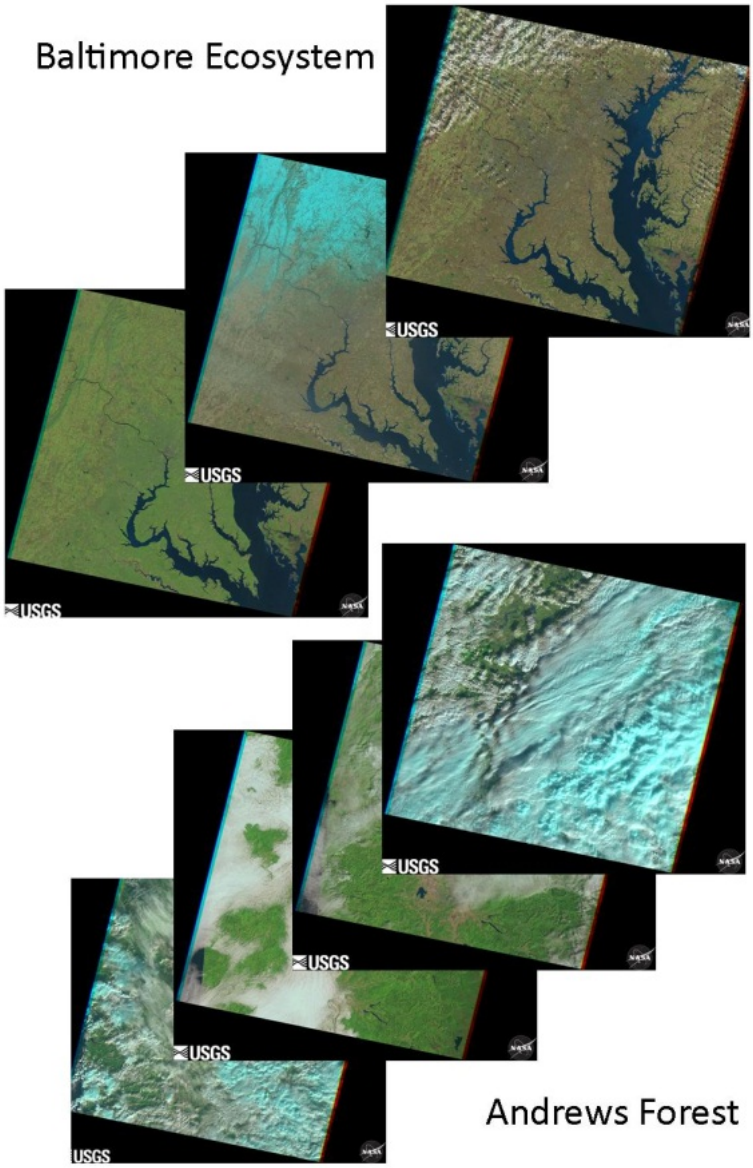
edit

Theresa Valentine (AND)

Work is currently being done to compile and make available a collection of Landsat scenes for each LTER site. The catalog will span from 1982 through 2011, and will provide 30 years of standardized historical data. Raw images will be available, as well as atmospherically corrected images. The corrected images will have two options: top of atmosphere reflectance, and surface reflectance. The geographic extent of the images per site include appropriate complete Landsat scenes; in other words, some LTER sites will have larger footprints because of their size and location.

The imagery will be available for download through PASTA. The catalog is expected to be completed by August of 2013.

For more information, contact Theresa Valentine at tvalentine@fs.fed.us or Theresa.valentine@oregonstate.edu.



LTER Controlled Vocabulary Workshop Planned

edit

John Porter (VCR)

In 2011, the LTER Controlled Vocabulary Working Group made substantial progress toward the goal of making LTER data available, efficiently and reliably, through searching or browsing. Milestones included creating a thesaurus of over 600 keywords (http://vocab.lternet.edu). With support from the LTER Network Office, the thesaurus was used to automatically enhance searches on the LTER Data Portal by automatically searching for synonyms and narrower terms. Data browsing was also enhanced by implementing a more complete browse interface. The group also produced a series of technical resources that can be used by LTER sites and researchers (http://im.lternet.edu/vocab_resources). These resources include tools for suggesting keywords (based on scanning documents), graphical viewers for the LTER Controlled Vocabulary, and a variety of web services that support searching, such as autocomplete-pull-down lists. The LTER Controlled Vocabulary was also fully integrated into EnvThes thesaurus being developed by the European ILTER, and PASTA is also incorporating improved search technologies that rely on the thesaurus.

On May 21 and 22, 2013 a new Controlled Vocabulary Working Group workshop will be held in Oyster, Virginia, with several goals, the most important of which will be to enhance the controlled vocabulary by adding new terms. To prepare for the workshop, several LTER Information Managers have been tasked with identifying new terms. Thus far, 203 additional terms have been proposed. They will be evaluated at the workshop and recommendations will be forwarded to the entire LTER Information Management Committee for adding terms. Eda Melendez, Jason Downing, and Don Henshaw queried the LTER sites for additional terms that sites would like to propose for addition to the LTER Controlled Vocabulary. Additionally, M. Gastil-Buhl has been mining the Metacat archive to identify additional terms that are widely used but not yet part of the vocabulary, while Duane Costa has provided a list of terms commonly entered by users to search the LTER Data Portal. In a related effort, Corinna Gries has been examining other sources for terms that were excluded from the original LTER Controlled Vocabulary, specifically for taxonomic terms and place names. The plan is to keep these separate from the LTER Controlled Vocabulary, because they may best be used at other places in the metadata than the keywords section. Margaret O'Brien, Kristin Vanderbilt, and Jason Downing have been working on providing definitions for each of the terms already in the Controlled Vocabulary. The plan is to create a semi-automated method of populating these definitions that will take advantage of existing web services for definitions, but also allow customization in the event that words used by ecologists differ from those in other areas of life.

Another goal of the workshop will be to get a group together to discuss possible additional uses for the LTER Controlled Vocabulary. It has already become clear that the LTER Controlled Vocabulary can have additional uses. Through its relationship with EnvThes, the vocabulary has been used to implement a prototype multilingual searching system (http://vocab.lternet.edu/ILTER). However, there may be other valuable uses for the vocabulary that extend beyond the finding of datasets, and this discussion will explore those options. In addition to Information Managers at the workshop, there will also be librarians concerned with meeting the requirements of the new NSF Data Management Plans, who may be able to provide additional insights regarding the construction and use of the LTER Controlled Vocabulary.

The IMC meets the PASTA Challenge

edit

John Chamblee (CWT), Margaret O'Brien (SBC)

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

In January of 2013, the LTER Network Office launched a production version of the new LTER data repository based on a framework and application suite known as PASTA (Provenance Aware Synthesis and Tracking Architecture). Its launch was a significant achievement by the LNO developers, and sites are on track to have their data into PASTA by the end of this year. This is excellent news, but there are still issues to resolve as we move forward. Here, we review some recent accomplishments and challenges associated with the ongoing rollout and implementation of PASTA.

As of this writing, there are 872 data packages from 12 sites loaded into the PASTA framework. Many sites are putting recent supplement funds to use adapting their EML-generating systems to meet PASTA's structural and data requirements. Others are thinking through their overall approaches to PASTA submission. In the process, many sites are re-evaluating their inventories, re-designing and/or integrating time series, and generally improving the quality of the data going into PASTA. These efforts will all contribute to greater overall availability for LTER data.

The accelerated timeline for PASTA's development also brings some significant challenges to both policy and site-level implementation. The adoption of PASTA as the basis of the LTER data cataloging system means that we will switch from the current metadata-only catalog to one designed for both data and metadata. Overall, this step will improve availability and accessibility because PASTA provides DOI-based identification for data packages, has the capacity for synthesis data provenance tracking, and more fully exploits the Ecological Metadata Language schema.

Some of the issues that involve the IMC are:

1. Now that PASTA is available and we are loading data into it, how will we (scientists and IMs) make the most of it? How will workflows be developed, advertised and used by our synthesis working groups? This will be a major topic at the upcoming IMC meeting in Fairbanks. Additionally, an IMC workshop will design documented workflows and provide practical, real-world experience that will inform best practices for EML metadata and PASTA development (Sheldon, et al., 2013).
2. NSF has stressed that LTER should have a "one-stop shop" for all data. The IMC will be providing feedback to the NIS web portal development process to ensure that LTER data are available to scientists in a way that makes them easily discoverable and accessible. In 2013, an IMC working group will enhance the LTER Controlled Vocabulary to further enable automatically enhanced searches of the catalog (Porter, et al., 2013).
3. Sites have always handled data that are sensitive or have restricted access (Network "Type II", http://www.lternet.edu/policies/data-access). In some cases, data distribution may even be prohibited by legally binding use or redistribution agreements (although it may be beneficial to post metadata). The PASTA framework requires that a data object be attached to every metadata record, and EML is fully capable of defining highly specific access rules. PASTA will be the basis of our one-stop catalog, but site-compatible practices for handling all levels of access are still being developed. The IMC will outline a full set of sites' catalog use cases and suggest potential solutions for handling all forms of sensitive, restricted, or provisional data at the Network level.
4. The current policy of collecting user information at the time of download continues to be a major topic of debate within the LTER Network, and several committees have outlined the costs and benefits of tracking data usage. The practice of collecting data usage information is linked to our data access policy, and so is the purview of the Science Council. Should changes in policy be enacted, the IMC will then determine their implications to our generation of data packages, and to the operational definition of "restricted data" (see previous point). We will also consider the implications of practices we develop for LTER as a DataONE node, since the access control policies we set will impact the ways in which DataONE displays our data.

We continue to operate in a period of rapid change. The good news is that we are making rapid progress. In many ways this is a watershed moment in LTER, when we are poised to capitalize on our long-term investments. We focused here on the challenges, but the fact remains that we should all be very proud of PASTA's existence as a production framework, and that IMC has pulled together to do what it takes to make it our "one-stop shop" for LTER data. As we move forward, the results of our collective effort will pay dividends in terms of both data availability and network-wide collaboration for years to come.

References:

Porter, et al., 2013. http://intranet.lternet.edu/content/enhancement-lter-controlled-vocabulary-support-data-synthesis

Sheldon, et al., 2013. http://intranet.lternet.edu/content/leveraging-pasta-and-eml-based-workflow-tools-lter-data-synthesis

LTER Data to be Part of Web of Knowledge

edit

James Brunt (LNO), Margaret O'Brien (SBC)

In October 2012, Thomson Reuters launched a component within their *Web of Knowledge* platform called the *Data Citation Index* (DCI), which will support discovery, access, citation, and attribution of digital data. Initially their work has focused on data sets deposited in repositories and which are used in the research published in the traditional scientific literature. Their system is planned to enable the discovery of data studies and data sets, and to help track the full impact of research and institutional output. After reviewing LTER data catalogs, Thomson Reuters has selected our Network as a key resource for this new initiative. Inclusion in the new Data Citation Index will provide LTER with increased visibility, data citation, and repository traffic, and additionally, to metrics of use similar to that provided for published papers.

The Thomson Reuters initiative has been discussed within IMExec, NISAC, and the Executive Board, and the LNO is acting as the point of communication for further work. LNO communicated a crosswalk between the DCI metadata and EML for analysis and scripted examples of accessing the PASTA API for Data and DOIs. We plan to encourage LTER scientists to cite data which will prime the pump for seeing LTER data cited elsewhere.

More information about the Data Citation Index can be found at http://wokinfo.com/products_tools/multidisciplinary/dci/

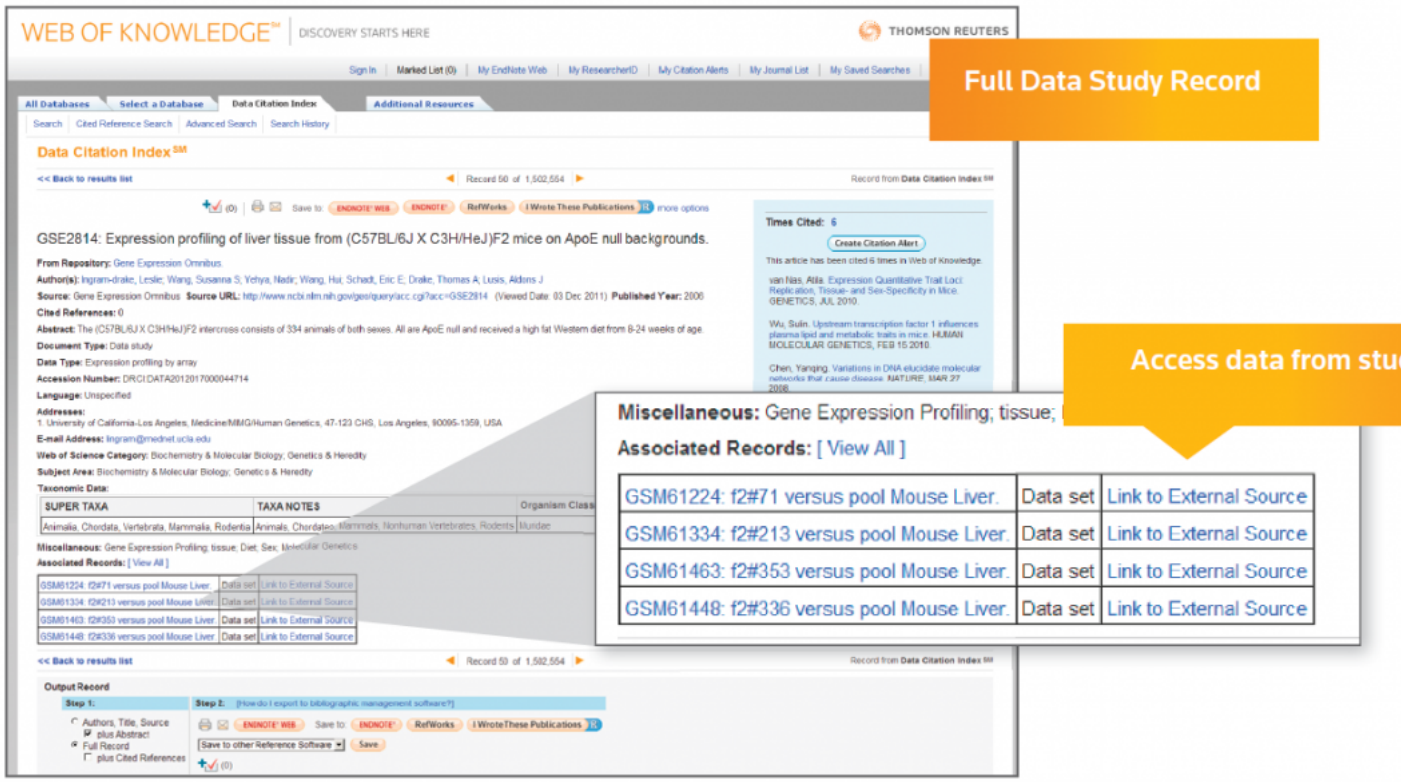


Figure 1 shows a screenshot of a DCI record.

Special thanks to Margaret O'Brien and 'Gastil' Gastil-Buhl for their contributions to this effort.

Good Tools And Programs

New DataONE Tools Help with Information Management Tasks

edit

John Porter (VCR)

The DataONE project (http://dataone.org) has recently come out with several tools and resources that LTER Information Managers and other researchers should be aware of as part of their "Investigator Toolkit" (http://www.dataone.org/investigator-toolkit). Here are some quick highlights:

- **DataUP** - a tool for reviewing spreadsheets to assure that best practices are followed. The tool comes in two forms, a plugin for Excel and a web application. Both provide reports on problems with spreadsheet data, such as inconsistent data types and missing entries. DataUP also supports metadata development and access to ONEShare - a data archive. See: http://dataup.cdlib.org/
- **DMP Tool** - a tool for preparing Data Management Plans. It provides guidance for a large number of federal and private funding agencies (including NSF Directorates), and often, information specific to your university. Plans can be entered into a web form, then exported to Microsoft Word or other text forms for inclusion in proposals. Old plans can be saved and modified for use with new proposals. See: https://dmp.cdlib.org/
- **OneR** - the DataOne R client lets you access ecological and environmental data from the DataONE network of repositories (which includes LTER) from inside the R statistical language. Data are accessed using the unique identifier associated with each dataset, which helps to make R code location- and user-independent. See: https://releases.dataone.org/online/dataone_r/

Additionally, DataONE has a library of **education modules** in the form of Powerpoint presentations that can be used to help train your researchers in the basics of ecoinformatics (http://www.dataone.org/education-modules). You can use and edit them as you would like because they are covered under a Creative Commons Zero license (no rights reserved). They also have a searchable **library of Best Practices** (http://www.dataone.org/best-practices) and a searchable **catalog of ecoinformatics tools** (http://www.dataone.org/software_tools_catalog).

Additional products are in the pipeline, so a periodic visit to http://dataONE.org is a good idea!

Using Python for Web Services and GIS Analysis

edit

John Porter (VCR)

Computer languages are a bit like cups of coffee - there is always another one not too far away! They also have their time. FORTRAN once ruled the roost, but now is used only for specialized applications. Similarly, C, PASCAL, Java and C++ all have had their time at the "top" as the "hot" language, and continue to be widely used today. So why do we need Python, a scripting language named after, not a snake, but a comedy troop (Monty Python)? And why is it such a "hot" language today? To answer those questions I'll talk a bit about the characteristics of Python and give some examples of how it is used at the Virginia Coastal Reserve LTER.

Python is an object-based, modular, scripting language with simple syntax. Taking those in order, each variable in Python is an object, and comes associated with methods/functions that can be used to manipulate it. Apart from basic functionality, Python uses "modules" that can be imported to support specific features (e.g., dates, database interfaces), thus reducing the potential "bloat" of trying to load all possible functions at one time. Python is a scripting language because one of its best uses is short programs in applications where blazing speed is not an issue. Python is not compiled, which makes it slower to run, but easier to write, with simplified debugging. Finally, gone are the declaration statements, curly braces, and semi-colons beloved by Java and C programmers. Instead, Python uses

white space for organization: line ends serve as statement terminators and indentation indicates the scope of loops. A major goal of Python design was to have programs that were easy to read, as well as to write... which is an important consideration in LTER, where long-term durability of software is often an issue.

One of the main reasons for using Python is that support for it is increasingly being built into 3rd party applications. For example, ESRI's ArcGIS software now supports the "ArcPy" library of functions, allowing you to run any of the tools built in to ArcGIS directly in Python. This allows automation of frequently performed tasks without requiring access to the ArcGIS graphical user interface. An additional benefit is that running a script in Python that uses ArcGIS functions is much faster than running those same functions in ArcGIS itself, which is a massive program. For example, one application I wrote that required 45 minutes to run as an ArcGIS Model, ran in only 22 minutes when using Python, presumably because much of the "overhead" associated with the ArcGIS interface was eliminated. Running ArcGIS tools directly in Python also aids with looping, because you can use the string manipulation functions in Python to create lists of files to be processed, then use a FOR loop to process each of them.

One nice feature of ArcGIS is that it will create basic Python programs for you. The ModelBuilder tool allows you to create "models" using a graphical interface (essentially a box and arrow diagram of a workflow). Once completed, you can "export" the model to a Python program. I frequently use this feature to develop and test the core functionality of a program, then edit the resulting Python script to add looping and other functions. One downside of using Python with ArcGIS is that there are still some rare, but annoying, cases where ArcPy will indicate that a function has completed when it is still running, thus causing file access errors with a subsequent statement to manipulate the file.

We also use Python in association with R to facilitate the import of some data logger files that have extensive header information that need to be "mined" for data (such as unit location, serial number, etc.). Parsing data out of headers is difficult-to-impossible to do in R, but because Python supports the use of "regular expressions", it is a powerful tool for parsing data out of complex (but regular) file structures. We use Python to pre-process the data logger files to collect information from the headers and prepend it to the actual data, thus creating a comma-separated-value (CSV) file where each line contains information from the header (e.g., the location where the instrument was deployed) along with the periodic readings (e.g., temperature, pressure). The CSV file is then ready for further processing using R.

A final example of Python use has been to work with some of the web services provided by PASTA. Its urllib2 module allows Python to act as a web browser to interact with web services, and it also supports several modules for manipulating XML documents. Here is a code snippet that queries PASTA to fetch an EML document and extracts the title and a list of entities:

```
# Import the needed modules
import urllib2
from datetime import datetime
import xml.etree.ElementTree as ET

# set the information for the dataset selection
pastaScope='knb-lter-vcr'
pastaId=25
pastaVersion=27
pastaFromTime='2012-01-01T00:00:00'
pastaToTime='2013-03-31T23:59:00'

# Set up an authentication string
# Note that you can then save userData string for future use.
uName='uid=ME,o=LTER,dc=ecoinformatics,dc=org'
pWord='mypassword'
userData="Basic " + (uName + ":" + pWord).encode("base64").rstrip()

# Get the data package metadata to extract dataset title and entities
# set the URL up for query and prepare a request
emlUrl="http://pasta.lternet.edu/package/metadata/eml/"+pastaScope+"/"+str(pastaId)+"/"+str(pastaVersion)
emlReq=urllib2.Request(emlUrl)
emlReq.add_header('Authorization', userData)
# execute the request and fetch the document
emlSock=urllib2.urlopen(emlReq,timeout=60)
emlString=emlSock.read()

# Parse the EML document returned using the ElementTree Module
emlRoot=ET.fromstring(emlString)

# print the title and search info, converting numbers to strings and contenating them together using +
print "Downloads of dataset "+pastaScope+"."+str(pastaId)+"."+str(pastaVersion)+ " between "+pastaFromTime+" and
"+pastaToTime+":\n"+emlRoot.find('./dataset/title').text

# use the 'findall' method of the emlRoot object to find all the entities, count them, then list them
entityRecords=emlRoot.findall('.//entityName')
print("contains "+str(len(entityRecords))+" data entities");
for entityRecord in entityRecords:
print(entityRecord.text)
```

The resulting output when run is:

```
Downloads of dataset knb-lter-vcr.25.27 between 2012-01-01T00:00:00 and 2013-03-31T23:59:00:
Hourly Meteorological Data for the Virginia Coast Reserve LTER 1989-present
contains 2 data entitiesVCR97018VCR_LTER_Hourly_Weather.zip
```

Some good Python tutorials (which feature interactive windows to allow you run exercises without actually installing Python on your system) are at: http://www.learnpython.org/ and http://www.codecademy.com/tracks/python/. If you'd prefer a free, full university-level course, try https://www.udacity.com/. Their Computer Science courses lean heavily on Python - and if you want, you can even pay to get the academic credit.

Good Reads

Big Data and the Future of Ecology

edit

Hope Humphries (NWT)

Hampton, S.E., C.A. Strasser, J.J. Tewksbury, W.K. Gram, A.E. Budden, A.L. Batcheller, C.S. Duke, and J.H. Porter. 2013. Big data and the future of ecology. Front Ecol Environ, 11(3):156-162, doi:10.1890/120103.

In an age of "big science" and "big data" (massive volumes of data not easily handled by the usual tools and practices), ecologists are encouraged in this article to be more proactive in contributing data, at all spatial and temporal scales, to publicly available data repositories. The case is made that the lack of a culture of data curation and sharing in ecology, in contrast to other scientific disciplines such as genetics, has resulted in a large pool of unavailable "dark data". The arguments for making such data publicly discoverable will be familiar to LTER information managers.

The authors provide a set of action items that individual ecologists could implement to increase their contribution to addressing large-scale environmental questions:

- Organize, document, and preserve data for posterity.
- Share data.
- Collaborate with networks of colleagues.
- Address data management issues with students and peers.

http://www.esajournals.org/doi/pdf/10.1890/120103

Emerging Computer Security Issues

edit

Aaron Stephenson (NTL)

Cisco has released their annual security report, which touches on a number of different topics. Of particular interest to information managers is the topic on users having multiple devices to access cloud services. Because of this new paradigm, data center traffic is expected to quadruple in next 4 years. Along with the increase in the number of devices and network connections comes an increase in the number of available attack vectors.

Another highlight from the report include is the changing nature of malware and spam. No longer do they originate from "suspect" sites that can be easliy blacklisted; instead, more and more they come from third party advertisements shown on legitimate mainstream websites, and through malicious links in email messages that spoof major brand names.

Traditional methods of securing endpoints and network perimeters will no longer be sufficient to guard against attack:

> *"attackers have become increasingly more sophisticated, going after the sites, tools, and applications that are least likely to be suspected, and users visit most frequently. Modern threats are capable of infecting mass audiences silently and effectively, not discriminating by industry, business, size or country. Cybercriminals are taking advantage of the rapidly expanding attack surface found in today's "any-to-any" world, where individuals are using any device to access their business network. As critical national infrastructure, businesses, and global financial markets continue their move to cloud-based services and mobile connectivity, an integrated, layered approach to security is needed"*

The report can be downloaded for free at Cisco's website:

http://www.cisco.com/en/US/prod/vpndevc/annual_security_report.html