LTER

LTER Databits

Information Management Newsletter of the Long Term Ecological Research Network

01001100 01010100 01000101 01010010

Fall 2013

Welcome to the Fall 2013 issue of Databits, which covers a variety of topics. A Commentary article gives an overview of network products, highlighting 2013 as a watershed year for the roll-out of community-developed IM tools. Several articles in this issue contain further information about these products. Two other commentaries extol the strengths of the EML Congruence Checker and offer a way to assess IM costs by source of support, an important consideration as changes in network operation are evaluated. Featured articles include use of the GCE Data Toolbox for automating sensor data harvesting and quality control; new DEIMS data discovery features (faceted searching and the Data Explorer); tools for securing data, a significant current topic; approaches for assuring data integrity; and the use of PASTA audit web services. Under "News Bits", an update is provided on the GeoNIS web service for processing spatial data and delivering mapping services to sites. The Good Reads section includes reviews of a book regarding "Big Data" and an article on common errors in ecological data.

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.

Hope Humphries (NWT) and **John Porter (VCR)**, Fall 2013 Co-Editors

Featured Articles

Commentary

News Bits

Good Reads

Featured Articles

Securing Your Data

edit

John Porter (VCR)

Recent headlines regarding the National Security Agency (NSA), have made it abundantly clear that “data security” is always a relative term. Does that mean we should just give up and throw open doors to our systems? Not so. The key is to make your system secure enough that malicious parties will simply find it too expensive or inconvenient to penetrate it. Often the key issues are not so much technical as personal. “Human engineering” is a key way hackers gain access to systems.

James Brunt has produced several excellent security-related blog entries at: http://lno.lternet.edu/blog/jbrunt. Notably “Passwords, passwords, passwords” details why passwords are both our “best friend” and our “worst enemy” when it comes to security. “On the Road: Fear and Living with Public Computing” discusses the perils of operating on open networks and “Protecting Your Digital Research Data and Documents” addresses backups. I’ll hit briefly on these topics, but I’d be remiss if I didn’t point you to these excellent contributions for more details.

The military term “defense-in-depth” is applicable to securing LTER systems. No single security solution is enough. It requires a series of tools to protect your data.

Outer defense: Computers should be protected by network firewalls that only allow only needed connections. Open Internet “ports” are an invitation to hackers seeking to penetrate your system. One colleague of mine installing a FTP server found that he had attempts to break into his system from overseas less than 45 minutes after the computer was first turned on. A firewall limiting access to specific services from specific computers or subnets can greatly limit the number of hackers banging away at your system.

Boundary defense: Long, complex (ideally random) passwords are the best defense against “brute force” attacks. However, they are also hell to generate, manage, and use, unless you use a password manager such as the open source “KeePass” software (http://keepass.info/). Such a manager lets you cut-andpaste passwords without having to type them at all. This makes it easier to use unique passwords (and if you are really tricky, unique login ids and security question answers) for each site you use. Randomly generated passwords are preferred because lists of millions of commonly-used passwords are available from hacker sites on the web. I showed a list of the top 25 passwords to a group of about 50 people – and the gasp from the audience was very audible. Now expand that list by a few orders of magnitude and you can see the problem. Also your system is only as protected as its least security-conscious user. Once someone can get in, the ability to do damage increases exponentially!

Use of unique passwords is a must for any system that contains anything of value, since every web site knows (or could know) the password you use there. Unscrupulous web sites that ingest passwords only to sell them on the open market are not unknown. Two-stage authentication, where a code is sent to your phone when you login from a new computer, is also a powerful way secure access and is increasingly available on commercial services.

Network encryption, including https, virtual private networks, and Secure Socket Layers (SSL) are required to keep your passwords from being intercepted in transit and should be used wherever they are supported.

Securing your electronic mail account is of special importance. This is because the password recovery functions of many web sites (including LTERnet) send you email that allows you to reset a forgotten password. Thus anyone who has access to your email account can “steal” your other accounts. Recently publicized incidents of identity theft or computer vandalism increasingly depend on a chain of actions – and often breaking into an email account is a key feature. Thus, encryption and a long, complex password are a must for your email account!

Internal Defense: Computers should be frequently updated to fix security holes that have been discovered in the operating systems and software. Computers need to be running up-to-date anti-virus software that detects viruses and malware – this is especially true for computers supporting electronic mail or uploading. Be cautious regarding the use of JAVA plugins in web browsers because JAVA-based exploits make up the vast majority of web malware. This is the reason for all the security messages have been popping up in browsers when JAVA is invoked.

Last Defense: Secure backups are the best way to prevent total disaster (e.g., total data loss). As discussed by James Brunt, backups (notice the plural – it’s important) need to be maintained at a variety of locations. Off-site (as far off-site as is feasible) backups provide the best protection. If you are using spinning disks as your backup medium over a network keep in mind that a hacker that penetrates your main computer system could also access, and delete or alter, your backup medium. So a copy that is not available online is highly recommended. Finally, a backup is really a backup only if it works. Using your backup system to periodically recover data from a backup is a must to assure that the system is operating properly. It also helps to be familiar with the restore procedure, so that when your data are at risk, you aren't needing to experiment.

Wrapup: Computer security, like the Red Queen in Alice in Wonderland, needs to keep running merely to stay in place. As quickly as security measures are adopted, people start working on how to circumvent them. Actually, often the reverse is true: new exploits demand new security responses. Thus computer security is always a relative, not absolute property. The key is to be at least a bit harder to penetrate than the next target!

Automating Data Harvests with the GCE Data Toolbox

edit

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

As described in the Spring 2013 issue of Databits, infusions of funding from the ARRA award to the LTER Network (Chamblee, 2013a) plus an NSF SI2 grant to Tony Fountain and colleagues (Gries, 2013) allowed us to make quantum leaps in both the functionality and usability of the GCE Data Toolbox for MATLAB software in 2012-2013. Accompanying funding for user training and support also allowed us to introduce more potential users to this software, and to work directly with new and existing users to take full advantage of this tool (Chamblee, 2013a; Henshaw and Gries, 2013; Peterson, 2013). This intensive work on the toolbox not only resulted in major improvements to the software, but allowed us to develop critical user support resources (https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox/wiki/Support) and establish an email list and Wiki pages to encourage ongoing peer support and collaboration. This process also provided the necessary momentum to remove remaining GCE-specific code from the main distribution and open the Subversion repository to public access, completing a 12-year transition from the toolbox being a proprietary GCE-LTER software tool to an open source community software package.

While the GCE Data Toolbox software can be leveraged by Information Managers in many ways to process, analyze, and distribute the data their LTER sites collect, the use case that has resonated most in training events is automating sensor data harvesting and quality control. The toolbox is ideally suited for automating data harvesting, because it combines prebuilt import filters for many common data logger formats (e.g. Campbell Scientific Instruments, Sea-Bird Electronics) and data services (e.g. NOAA HADS, NOAA NCDC, USGS NWIS, Data Turbine) with editable metadata templates containing attribute descriptors and QA/QC rules. By applying a metadata template on import, users can simultaneously parse, document, and quality control raw data in a single workflow step (figure 1).

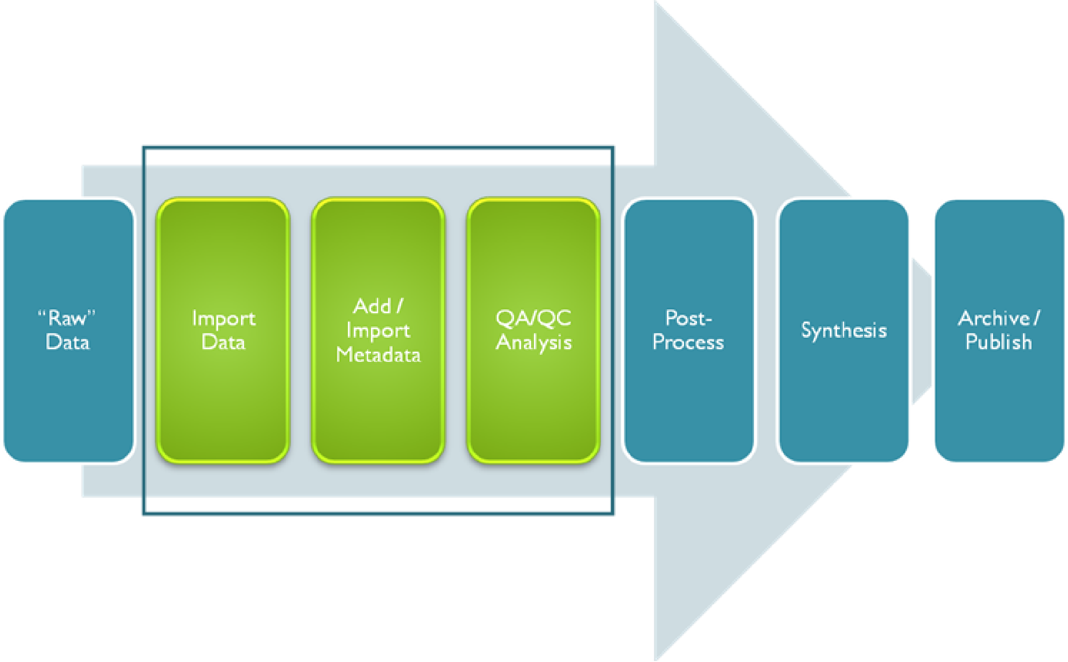


Figure 1. Coupling import filters with metadata templates containing pre-defined QA/QC rules allows the GCE Data Toolbox to accomplish three major stages of the data management cycle in one step.

Metadata content and QA/QC rules can be defined, copied, imported and even synchronized across multiple templates using form-based GUI applications, greatly simplifying management and re-use of this critical and time-intensive information. Tools and demo workflows are provided for generating basic data distribution websites, complete with indexed data summary pages, plots and downloadable data files (e.g. see http://gce-lter.marsci.uga.edu/public/file_pickup/toolbox_demo/data/index.html). These pages can easily be "skinned" using XSLT and CSS to match the appearance of an LTER site's web page for production use. Furthermore, once a data harvesting workflow is configured it can be executed on a timed basis using MATLAB's built-in timer objects, automatically appending newly acquired data to existing data sets to produce long-term time series data sets (figure 2). The GCE Data Toolbox therefore provides users with a comprehensive yet practical out-of-the box solution for many real-time sensor data harvesting scenarios (Chamblee, 2013b). Six LTER sites have already implemented harvesters using this software (GCE, CWT, AND, NWT, NTL, and HBR), and several other sites plan to as well. We are currently compiling a set of documented use cases for inclusion in an upcoming LTER Network News article to encourage further adoption.

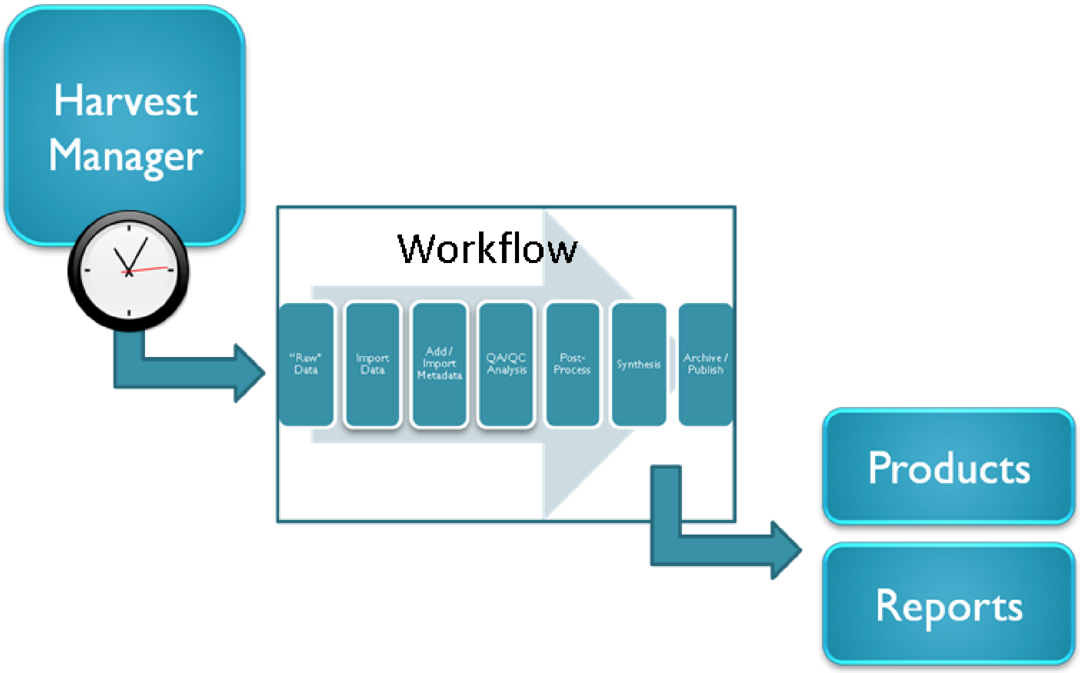
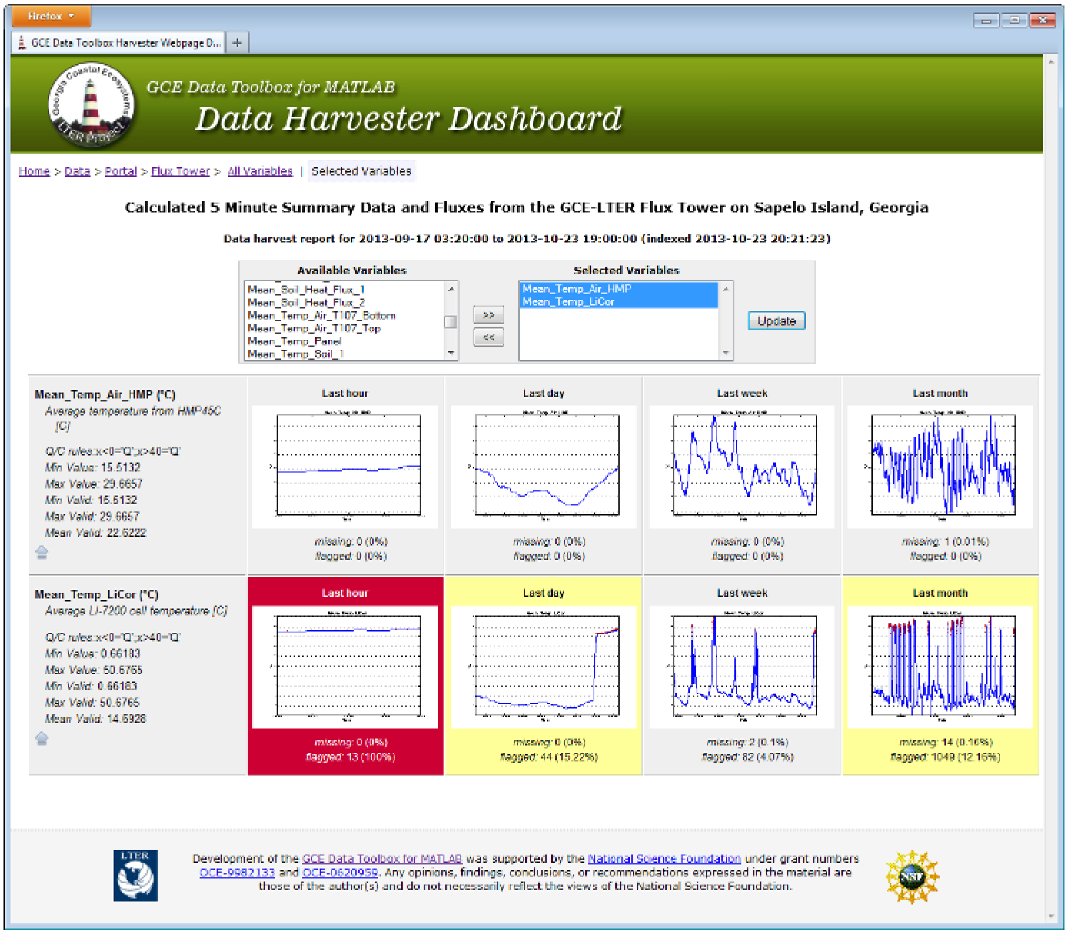


Figure 2. Data harvesting workflows can be executed on a schedule by registering them with the GCE Data Toolbox harvest manager. Harvests can then be queried, started or stopped using the toolbox GUI application, and continue to run in the background as long as MATLAB is running.

Although the supplemental funding has now ended, the GCE and CWT LTER sites continue to collaborate on development of the toolbox as part of core IM activities, with input from the broader toolbox user community. Recent work includes developing a fully automated web dashboard application for monitoring the status of real-time data harvesting systems (figure 3), including emailing of quality reports when problems arise, and completing support for registering metadata content in the Metabase Metadata Management System for archiving data in PASTA. Support for generating EML metadata directly from the toolbox is also nearly complete. We will also seek additional funds next year to continue development on the toolbox software, and continue to look for opportunities to lead or participate in hands-on training events to engage with more potential user groups.



GCE Data Toolbox for MATLAB

Data Harvester Dashboard

Home > Data > Portal > Flux Tower > All Variables | Selected Variables

Calculated 5 Minute Summary Data and Fluxes from the GCE-LTER Flux Tower on Sapelo Island, Georgia

Data harvest report for 2013-09-17 03:20:00 to 2013-10-23 19:00:00 (indexed 2013-10-23 20:21:23)

Available Variables	Selected Variables
Mean_Soil_Heat_Flux_1 Mean_Soil_Heat_Flux_2 Mean_Temp_Air_T107_Bottom Mean_Temp_Air_T107_Top Mean_Temp_Panel Mean_Temp_Soil_1	Mean_Temp_Air_HMP Mean_Temp_LiCor

	Last hour	Last day	Last week	Last month
Mean_Temp_Air_HMP (°C) *Average temperature from HMP45C [C]* *Q/C rules:* x<0='Q';x>40='Q' *Min Value:* 15.5132 *Max Value:* 29.6657 *Min Valid:* 15.5132 *Max Valid:* 29.6657 *Mean Valid:* 22.6222	*missing:* 0 (0%) *flagged:* 0 (0%)	*missing:* 0 (0%) *flagged:* 0 (0%)	*missing:* 0 (0%) *flagged:* 0 (0%)	*missing:* 1 (0.01%) *flagged:* 0 (0%)
Mean_Temp_LiCor (°C) *Average LI-7200 cell temperature [C]* *Q/C rules:* x<0='Q';x>40='Q' *Min Value:* 0.66183 *Max Value:* 50.6765 *Min Valid:* 0.66183 *Max Valid:* 50.6765 *Mean Valid:* 14.6928	*missing:* 0 (0%) *flagged:* 13 (100%)	*missing:* 0 (0%) *flagged:* 44 (15.22%)	*missing:* 2 (0.1%) *flagged:* 82 (4.07%)	*missing:* 14 (0.16%) *flagged:* 1049 (12.16%)

Development of the GCE Data Toolbox for MATLAB was supported by the National Science Foundation under grant numbers OCE-9982133 and OCE-0620959. Any opinions, findings, conclusions, or recommendations expressed in the material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Figure 3. Automated data harvest dashboard webpage generated by the GCE Data Toolbox. Note that multiple views are supported using XSLT stylesheets, including selected variables (shown) and all variables displayed.

Literature Cited

Chamblee, J. 2013a. GCE and CWT Host Successful Workshop to Demonstrate, Improve, and Promote the Adoption of the GCE Data Toolbox for Matlab. LTER Databits - Information Management Newsletter of the Long Term Ecological Research Network, Spring, 2013 issue. (http://databits.lternet.edu/spring-2013/gce-and-cwt-host-successful-workshop-demonstrate-improve-and-promote-adoption-gce-data-t)

Chamblee, J. 2013b. Coweeta LTER Upgrades Sensor Stations by Implementing the GCE Data Toolbox for Matlab to Stream Data. LTER Databits - Information Management Newsletter of the Long Term Ecological Research Network, Spring, 2013 issue. (http://databits.lternet.edu/spring-2013/coweeta-lter-upgrades-sensor-stations-implementing-gce-data-toolbox-matlab-stream-data)

Gries, C. 2013. Integrating Open Source Data Turbine with the GCE Data Toolbox for MATLAB. LTER Databits - Information Management Newsletter of the Long Term Ecological Research Network, Spring, 2013 issue. (http://databits.lternet.edu/spring-2013/integrating-open-source-data-turbine-gce-data-toolbox-matlab)

Henshaw, D. and Gries, C. 2013. Sensor Networks Training Conducted at LNO. LTER Databits - Information Management Newsletter of the Long Term Ecological Research Network, Spring, 2013 issue. (http://databits.lternet.edu/spring-2013/sensor-networks-training-conducted-lno)

Peterson, F. 2013. My Experiences as a Participant in the Sensor Training Workshop. LTER Databits - Information Management Newsletter of the Long Term Ecological Research Network, Spring, 2013 issue. (http://databits.lternet.edu/spring-2013/my-experiences-participant-sensor-training-workshop)

Using PASTA Audit Web Services

edit

John Porter (VCR)

One of the ways to motivate investigators to share their data is to show them that shared data is serving a scientific purpose. No one wants to spend the time and effort to produce data resources that are never accessed (a write-only dataset). Additionally, some investigators want information regarding who is accessing their data. One solution to meeting these information needs is to publish periodic reports on downloads and to distribute these reports to the contacts or authors for each dataset.

The PASTA data portal (https://portal.lternet.edu) provides access to Audit services under its “Tools” menu. For example, the “Data Package Access Report” allows searches for information regarding downloads of specific components of data packages (entire packages, metadata, quality reports and specific entities) filtered by dates, users and user groups. As such they are a valuable tool if you wish to assess resource usage for annual reports or other purposes. However, the pages produced are not ideal for communicating to investigators information on how the data they provided are being used. Issues for investigators include:

1. you need to be authenticated (login) to view audit pages and many investigators don’t remember their LTERnet passwords;
2. audit reports need to be specifically requested, so they can't be automatically provided on a periodic basis;
3. listings tend to be in a lengthy log format, reporting the date and time of each individual download of a resource; and
4. although you can search for all the downloads within a given scope (e.g., knb-lter-vcr), there is no mechanism for retrieving only packages related to a particular investigator. Therefore, multiple searches are required to retrieve all of the packages related to a researcher.

Fortunately, all of these issues can be addressed by using the underlying web services that are used “behind the scenes” in the web portal.

PASTA supports a wide array of audit web services (see: https://pasta.lternet.edu/audit/docs/api for full documentation) that allow programs, as well as users, to access the information on PASTA usage. Invoking a web service returns an XML file containing the requested information. Programs can easily parse the returned information to create individualized summaries that can be shared with investigators.

There are some “tricks of the trade” when writing programs to access PASTA. Examples given are from a Python program written to provide reports to VCR/LTER investigators and is available at: https://svn.lternet.edu/websvn/listing.php?repname=VCR&path=%2Ftrunk%2FPASTAsummary

- **PASTA Basics** – You need to understand the basics of how PASTA identifies particular resources, in particular that each package has a **scope** (e.g., knb-lter-vcr), an **identifier** (e.g., 25) and a **revision** (e.g., 29) – typically strung together with slashes, so that knb-lter-vcr/25/29 specifies a unique package. As discussed below, entities (data files) associated with a particular package have resourceIds that can be appended to the packageId to refer to an individual entity – but these are relatively cryptic and not something you’d ever want to type in yourself.
- **Authentication** – Just as the Portal requires a login to access audit reports, so do the web services. Most programming languages support access to web services using variants of the CURL library. To avoid the need to encode passwords, it is possible to generate a “userData” string that encodes the username and password (here using Python):

```
uName='uid=VCR,o=LTER,dc=ecoinformatics,dc=org'
pWord='myPassword'
userData="Basic " + (uName +":" + pWord).encode("base64").rstrip()
```

The userData string can be used to add an “Authorization” header to a request (here with the Python version of CURL [urllib or urllib2]):

```
req.add_header('Authorization', userData)
```

- **Selecting Resources** – Each resource in PASTA has a unique identifier (resourceId) that needs to be used to access it. Thus accessing audit records regarding a given data entity (i.e., data file) requires that you first use the normal PASTA search services to find the appropriate entity (e.g., http://pasta.lternet.edu/package/data/eml/knb-lter-vcr/25/29 lists the resourceIds for the two entities associated with that package: ed6cf6c6df81ce0de14caf429aef63ef and c5b325e8182f30350782fb06be53be7c). These can then be used to access the relevant audit log entries for those entities.

As noted above, PASTA returns results as easily-parsed XML. There are such a wide variety of XML processing systems, I won’t go into the details regarding how that information is extracted, but anything, including a simple string search, can be used.

In the sample program XML is used both for retrieving data and for output. The PASTAsummary.py program burrows down into the package structure and even the original EML metadata to extract the title of the dataset, its list of contacts, and the entities it contains. It then queries the PASTA audit services to produce an XML structure containing the summaries. That XML can then be converted using a stylesheet into a customized HTML display (see figure) that can be saved or emailed. Sample output from PASTAsummary.py program

The utility of the web services in PASTA demonstrate that this is an approach that could more widely be applied to LTER Network systems, exposing capabilities while not restricting formats of outputs.

Data Integrity, The Phrase I Don't Hear Enough

edit

Ryan Raub (CAP)

Data integrity is basically the maintenance and assurance of the accuracy and consistency of data throughout its life-cycle. This aspect of data management is a key component of good archival system practices, and can have frightening consequences if omitted.

One of the scariest scenarios for an data archival system is corruption going un-noticed for a long period of time, perhaps longer than your backup recovery window. Imagine finding out that someone three years ago inadvertently opened up a published data file and changed one value. You have now unintentionally distributed this erroneous data to an unknowable quantity of users over the last three years. This is what we want to prevent, by adding assurances to the data you archive so that you can be sure it remains unchanged.

Data corruption could even occur with the file or operating system and it's up to your archival practices to discover any changes. You shouldn’t just rely on the file system for preservation of archival data; file systems are not perfect and do have a small percentage of problems. With the increasing volume of data, the occurrence of these problems increases proportionally. There are some file types that have some redundancy or checksums built into their format (e.g. zip or tar), however these redundancy features are only intended to answer the question "is this a valid file?" and not "has this file changed?".

A very common way of checking the contents of a file (or folder) for changes without duplicating the data is to compare a computed hash value (or checksum) of the data with a prior known hash value. There are several standard options for hashing functions with pros and cons, but for the scope of this article I’m going to recommend the SHA-1 hashing algorithm. These hashing functions will always return a fixed sized output (e.g. 40 character string) given any input file, regardless of the file size. Additionally these functions will be able to compute different outputs given two files of any size that only differ by one bit. So even the smallest change in a terabyte sized file is noticeable.

If you want to start data integrity checks for a small volume of files (less than one gigabyte) you can easily use a version control system like Git to store prior versions of files and their hashes. Git is a powerful tool and I would recommend it to anyone who wants a simple way to “keep tabs” on a directory (read more). The less than one gigabyte per directory (repository) is just a rule of thumb; it doesn’t have any hard limits. However if you are beyond that, there are probably better ways to achieve this goal.

If you cannot split your data holdings into less than one gigabyte directories, you can use some simple command-line tools to create a list of the files and their checksums (commonly referred to as a manifest). I’ve put together a simple linux bash script to generate and compare a list of checksums for a directory. This won't tell you what has changed in a file, only that the file has changed.

As the data volumes grow, you'll need to scale your tools accordingly. However the principles remain the same: compute a hash of a file periodically and see if the hash has changed. You can even use hashing functions in databases to catch changes within tables.

Other data archival systems like the Planetary Data System (PDS) that NASA uses have standards that require the hash to be stored in the metadata for each dataset. They even go as far as to require data integrity checks to be run over their entire data holdings on a monthly basis to ensure that nothing gets altered. Granted, the PDS operates at a much larger scale compared to the LTER, but the goals are the same. Perhaps we should consider adopting storage of the data file hash values as part of our data management best practices.

Opening the Data Vault with the Drupal Ecological Information Management System

edit

Inigo San Gil (MCM), Kristin Vanderbilt (SEV), Corinna Gries (NTL), Jeanine McGann (UNM), Marshall White (LNO), Eda Melendez-Colom (LUQ), Aaron Stephenson (NTL), Jim Laundre (ARC), Hap Garritt (PIE), ken Ramsey (JRN), Philip Tarrant (CAP), Ryan Raub (CAP), David Julian (CAP), Chau Chin Lin (TFRI), David Blankman (ILTER), Atzimba Lopez (Mex-LTER), Cristina Takacs-Vesbach (MCM), Palantir.net and Dave Reid.

This article covers two innovative features of the Drupal Ecological Information Management System (DEIMS): Faceted Searches and the Data Explorer.

Introduction

LTER sites have been charged with managing and archiving their data for long-term public use since the inception of the NSF LTER program. Due to the fact that LTER research is site based, long-term observations are highly optimized for a particular ecosystem and most datasets become re-usable only when collection methods and sampling contexts are well documented. Hence, a large variety of datasets are currently curated at LTER sites and a search for particular data can become a search for the proverbial needle in a haystack. Availability of data is no longer the problem with more than 6000 datasets published by the LTERs; the problem now is optimizing the data search. DEIMS provides several approaches to a more successful data search and tools for initial data exploration.

DEIMS is a powerful tool for managing most information products associated with an LTER, field station, or research lab. Significantly, it includes a web-based metadata editor for describing datasets and a module for generating EML, BDP and ISO compliant metadata. In addition, DEIMS includes web-based publication management, research project information, people associated with the site, images, and other information as needed. All of these interlinked resources benefit from secondary relations formed using common keywords that a DEIMS allowed user selects from either the LTER controlled vocabulary (Porter, 2013), the LTER Core Areas vocabulary, or other site-centric keyword families. The DEIMS site information management team completes the information curation process using the DEIMS *workbench*. Search results are displayed in DEIMS in various ways using Drupal's (Dries, powerful ability to create custom listing pages (or views) of data.

Thus, based on Drupal's (Dries, 2001) inherent capability of linking of different information concepts, a data consumer can find datasets produced by a person directly on the person profile page, or linked to a research site on the page describing that site, or linked to a specific research project on the page illustrating that project, or in connection with a journal publication. In addition to these different ways of accessing information on a typical DEIMS site, a faceted search has been implemented in the latest DEIMS version.

DEIMS Data Discovery

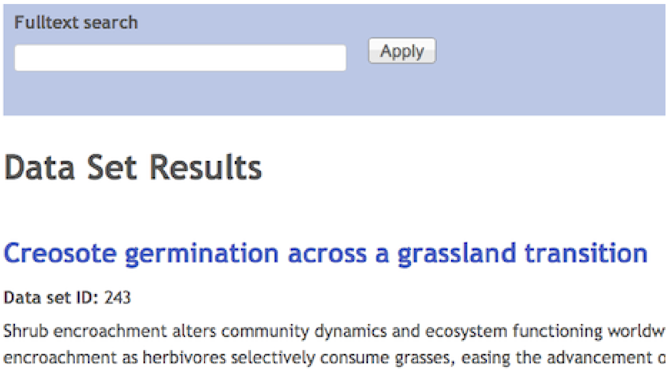
In past Databits issues (San Gil, Spring 2013 and Fall 2011; Gries et al., Spring 2010) we have written about DEIMS' approach to data discovery. Instead of a single Google-like interface to data discovery, DEIMS offers a variety of pathways to discover data (San Gil et al., 2010). A data consumer can search for data by a person's name (from the person profile page or data catalog interface), by association with a journal publication, or by particular location set or temporal range. Now, there is a new and even more specific DEIMS feature for the LTER data consumer -- faceted searches, the ability to narrow initial results from a broad search.

What is a faceted search?

Wikipedia defines faceted search as (Tunkelang, 2009): 'Faceted search, also called faceted navigation or faceted browsing, is a technique for accessing information organized according to a classification system, allowing users to explore a collection of information by applying multiple filters. A faceted classification system classifies each information element along multiple explicit dimensions, enabling the classifications to be accessed and ordered in multiple ways rather than in a single, pre-determined, taxonomic order.' Hence, faceted searching allows us to refine initial, broad, Google-like searches. It consists of the grouping of search results under narrow sub-categories. This narrowing helps find the appropriate set of results among hundreds of relevant results from an initial search. Most people have used a similar feature in many websites, such as Google shopping or at a variety of retail websites, i.e., an initial search for a 'refrigerator' yield a long list of results that can be narrowed further selecting price, brand, color, and a number of other filtering options, making the search experience better.

DEIMS has a number of facets or filters to improve your search experience as well. You can narrow the search for data sets by the lead principal investigator, by location, by duration, or through the tags used to classify the dataset.

The screenshot below shows the initial search interface for datasets -- by default a user sees the Google-like search box, and a constantly adapting listing of results matching the current search criteria:



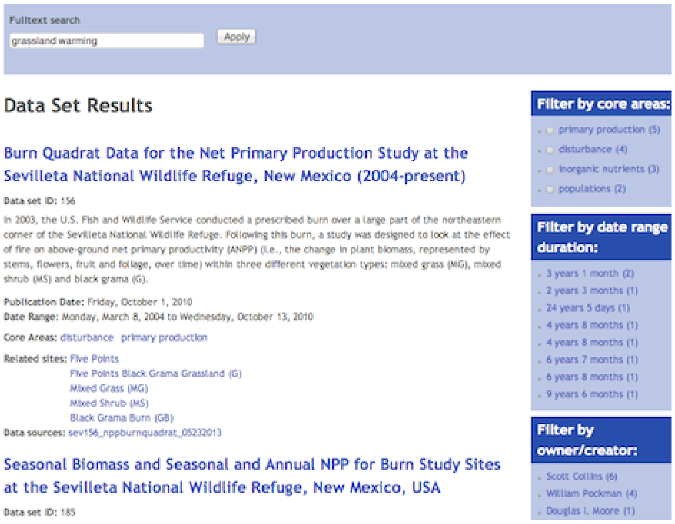
Data Set Results

Creosote germination across a grassland transition

Data set ID: 243

Shrub encroachment alters community dynamics and ecosystem functioning worldw
encroachment as herbivores selectively consume grasses, easing the advancement o

For example, a data consumer may start his/her initial search using the box, perhaps narrowing the result set to a fraction of all the data set catalog holdings. As an example, let's look at the initial results set using the search term "grassland warming". The text search box provided by DEIMS implements most comon simple searches, e.g., this would search for 'grassland' or 'warming' and provide a result set based on the index ranking from the combined terms. DEIMS adds the faceted search power: In the screenshot below all applicable facets that narrow the results further appear in the blocks on the right hand side:



Fulltext search

grassland warming [Apply]

Data Set Results

Burn Quadrat Data for the Net Primary Production Study at the Sevilleta National Wildlife Refuge, New Mexico (2004-present)

Data set ID: 156

In 2003, the U.S. Fish and Wildlife Service conducted a prescribed burn over a large part of the northeastern corner of the Sevilleta National Wildlife Refuge. Following this burn, a study was designed to look at the effect of fire on above-ground net primary productivity (ANPP) (i.e., the change in plant biomass, represented by stems, flowers, fruit and foliage, over time) within three different vegetation types: mixed grass (MG), mixed shrub (MS) and black grama (G).

Publication Date: Friday, October 1, 2010
Date Range: Monday, March 8, 2004 to Wednesday, October 13, 2010

Core Areas: disturbance primary production

Related sites: Five Points
Five Points Black Grama Grassland (G)
Mixed Grass (MG)
Mixed Shrub (MS)
Black Grama Burn (GB)
Data sources: sev156_nppburnquadrat_05232013

Seasonal Biomass and Seasonal and Annual NPP for Burn Study Sites at the Sevilleta National Wildlife Refuge, New Mexico, USA

Data set ID: 185

Filter by core areas:

- □ primary production (5)
- □ disturbance (4)
- □ inorganic nutrients (3)
- □ populations (2)

Filter by date range duration:

- 3 years 1 month (2)
- 2 years 3 months (1)
- 24 years 5 days (1)
- 4 years 8 months (1)
- 4 years 8 months (1)
- 6 years 7 months (1)
- 6 years 8 months (1)
- 9 years 6 months (1)

Filter by owner/creator:

- Scott Collins (6)
- William Pockman (4)
- Douglas I. Moore (1)

Data consumers can further filter the results by choosing one of the many data set owners, or by selecting a dataset timespan (duration). Another result-narrowing facet is the site-specific thematic keywords. A useful narrowing criteria is provided by the LTER Core Areas facet, the first block on the right in the figure above. If a data user browes the above page using a mobile device, these blocks would be rendered at the bottom of the page instead of to the side. Data facets are powered by default by the DEIMS relational database indexes. However, the data facets can be configured to consume an Apache Solr (Apache Foundation, 2010) created index. Data set facets can be extended with a variety of existing widgets, including graphs, slides, and tagclouds. The DEIMS facet search feature extends the Drupal facet API contributed module in conjunction with the context contributed module. In addition to documentation provided in the respective module pages, the documentation pages for the extension explain how to extend DEIMS out-of-the-box capabilities. Further set-up documentation and help on how to configure the faceted search is accessible through the DEIMS project page.

Once a dataset of interest has been found, a full data package is offered to the data consumer. However, often the data consumer is only interested in a subset of the data. DEIMS offers the data consumer the ability to explore the data and further narrow, dissect, and subset the original package.

Exploring LTER data with DEIMS - the Data Explorer

What is the Data Explorer?

The Data Explorer is a DEIMS feature that enables a user to connect to a relational database to expose and query its data holdings.

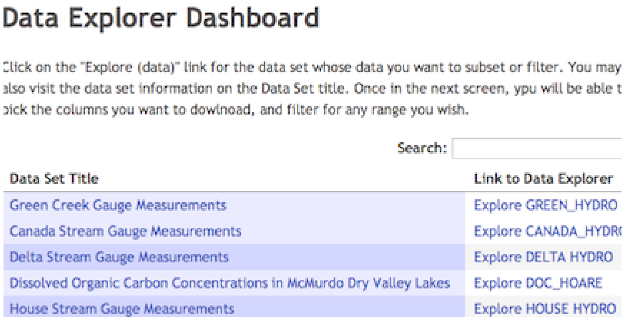
How do we use the Data Explorer?

The Data Explorer (DE) creates a query system for relational databases in DEIMS. When the DE module is configured, a data consumer may access a query page for each of the datasets catalogued and described in DEIMS. Each DE query page has two parts. The top part allows the data consumer to select which fields he/she wants to query, effectively subsetting a data package. The lower part of the query page allows filtering value ranges or thresholds from each of the data table columns.

For example, a hydrological table may have a *date* column, a *temperature* column, a *discharge* column, *conductivity*, and *quality flag* columns. The data consumer may just want to download the date and discharge values, which is done using the top part of the DE interface. In addition, the data consumer may want to narrow the data to a range of dates, set a temperature threshold, or exclude rows that have been tagged with quality control codes qualifying the validity of the overall measurements. This is done in the lower part of the data explorer. After submitting the query, the data consumer can either preview the results, or download them as a comma delimited file.

The next series of examples will illustrate this process using data collected by the McMurdo Dry Valleys Stream Team at the Delta Stream, led by Diane McKnight. The particular DEIMS setup may vary a bit depending on look and feel customizations or other improvements, but should essentially resemble what is shown here.

Starting with the DE Dashboard, a page that is accessed through the main menu (this example URL is the DEIMS default location http://example.com/data-explorer-dashboard):



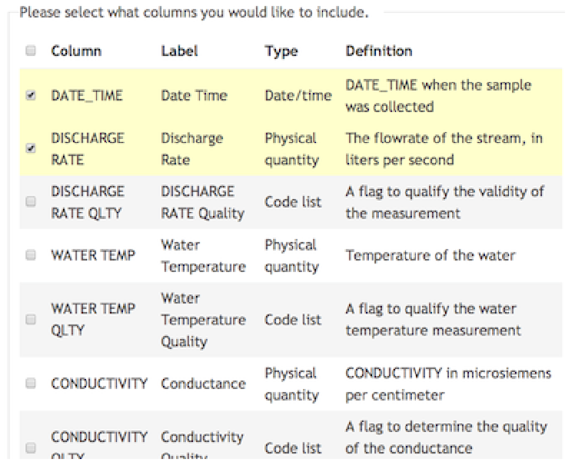
Data Explorer Dashboard

Click on the "Explore (data)" link for the data set whose data you want to subset or filter. You may also visit the data set information on the Data Set title. Once in the next screen, ypu will be able t pick the columns you want to dowlnoad, and filter for any range you wish.

Search:

Data Set Title	Link to Data Explorer
Green Creek Gauge Measurements	Explore GREEN_HYDRO
Canada Stream Gauge Measurements	Explore CANADA_HYDR(
Delta Stream Gauge Measurements	Explore DELTA HYDRO
Dissolved Organic Carbon Concentrations in McMurdo Dry Valley Lakes	Explore DOC_HOARE
House Stream Gauge Measurements	Explore HOUSE HYDRO

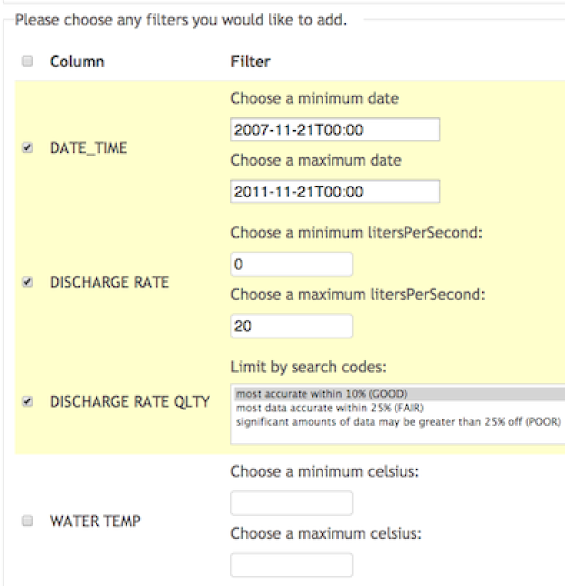
Data exploration is initiated clicking on the link “Explore DELTA HYDRO” in the right-hand column; the link leads to the Data Explorer query page, which is divided into two parts. The upper part allows us to select which of the columns we want to query and download data from.



Please select what columns you would like to include.

□	Column	Label	Type	Definition
☑	DATE_TIME	Date Time	Date/time	DATE_TIME when the sample was collected
☑	DISCHARGE RATE	Discharge Rate	Physical quantity	The flowrate of the stream, in liters per second
□	DISCHARGE RATE QLTY	DISCHARGE RATE Quality	Code list	A flag to qualify the validity of the measurement
□	WATER TEMP	Water Temperature	Physical quantity	Temperature of the water
□	WATER TEMP QLTY	Water Temperature Quality	Code list	A flag to qualify the water temperature measurement
□	CONDUCTIVITY	Conductance	Physical quantity	CONDUCTIVITY in microsiemens per centimeter
□	CONDUCTIVITY QLTY	Conductivity Quality	Code list	A flag to determine the quality of the conductance

Note that only the first two choices are checked because in this example we are only interested in the date that the discharge (stream flowrate) was measured and the discharge value itself. The rest of the data columns remain unchecked. Furthermore, in this example, we want to filter by a range of values. The lower part of the DEIMS DE query page allows us to create customized ranges to narrow the final results even further:



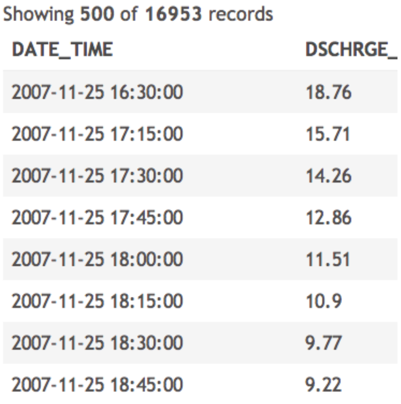
Please choose any filters you would like to add.

□	Column	Filter
☑	DATE_TIME	Choose a minimum date 2007-11-21T00:00 Choose a maximum date 2011-11-21T00:00
☑	DISCHARGE RATE	Choose a minimum litersPerSecond: 0 Choose a maximum litersPerSecond: 20
☑	DISCHARGE RATE QLTY	Limit by search codes: most accurate within 10% (GOOD) most data accurate within 25% (FAIR) significant amounts of data may be greater than 25% off (POOR)
□	WATER TEMP	Choose a minimum celsius: Choose a maximum celsius:

Using the datetime range filter, we have selected four years' worth of data -- from November 11, 2007 to the same day in 2011. We also filtered for high-confidence values (tagged as "GOOD": most accurate within 10%) and limited the results to discharge values at or below twenty liters per second. We could have added any other ranges to the rest of the column filters. It is important to use the checkbox to the left of the filter we want to apply.

Finally, there are the "Web Preview" and "Download" buttons at the bottom of the query page. Be patient with the "Download" for large datasets. Depending on the query, you may be requesting gigabytes worth of data!

To conclude this example, here is what a "Web Preview" of the results would look like for the example we are illustrating:



Showing **500** of **16953** records

DATE_TIME	DSCHRGE_
2007-11-25 16:30:00	18.76
2007-11-25 17:15:00	15.71
2007-11-25 17:30:00	14.26
2007-11-25 17:45:00	12.86
2007-11-25 18:00:00	11.51
2007-11-25 18:15:00	10.9
2007-11-25 18:30:00	9.77
2007-11-25 18:45:00	9.22

DEIMS DE requires adopters to transfer file-based data into a relational database. Currently DEIMS DE connects with Oracle's MySQL, the MySQL fork MariaDB, and Postgres, but it can be extended to connect to Oracle R-series databases and other flavors (SQLite, Microsoft SQL server, etc). Originally, the DEIMS group wanted to task the contractors with the development of a similar query directly against comma delimited files and spreadsheets; however, funding constraints forced the group to limit the DE feature to relational databases. We will pursue funding to extend these features. Documentation on how to configure the DE is available at the DEIMS project pages. You can always contact the authors of this article for help or assistance in this or any other DEIMS matter.

The DEIMS Data Explorer module is inspired by the original North Temperate Lakes LTER Data Catalog, a custom development led by Barbara Benson and transferred to the current NTL DEIMS Data Catalog (based on Drupal 6, and developed by Preston Alexander and others). Supplemental funding to NTL was allocated to the DEIMS project specifically to expand the original DEIMS Data Catalog module. The main developer, Dave Reid, was tasked with the generalization of the module with the goal of making the DE accessible to any LTER site and beyond.

Concluding remarks and next steps

In this article, we have covered the DEIMS grassroots efforts in expanding access to LTER data vaults. DEIMS exposes data holdings using state-of-the-art discovery methods, including faceted searches. Another feature we have covered is the ability to perform data subsetting, avoiding unnecessary downloads of massive datasets. DEIMS exports the metadata and data contents using the LTER-adopted metadata specification -- the Ecological Metadata Language, with PASTA-ready compliance. In addition, DEIMS also offers its metadata holdings using the Biological Data Profile specification, a profile of the content standard for digital geospatial metadata used by the US federal agencies. International visitors may use DEIMS' ability to export metadata formatted using the International Standards Organization (ISO) standards 19115, 19109 and 19110, expressed as an extension of the ISO standard 19139 XML implementation.

DEIMS development is ongoing. We are in the midst of adding charts and graphs to the DE outputs, an effort conducted in collaboration with our international partners at ILTER, specifically the IM committee chaired by David Blankman. Also, we will produce a road map for indexing metadata using Apache Solr, a technology already in use by Tai-Bif (Shao et al, 2013) and DEIMS instances at the Taiwan Forestry Research Institute. We also are exploring adding more community-developed widgets and styles to the DEIMS search facets.

We will seek funding to foster the growing DEIMS community, an essential component to sustain the current momentum of standardadization within LTER sites. Specifically, we need to nurture DEIMS training in both the use and management of facets. It has been a relative long time since we have conducted training. We also need funding to continue developing the system and addressing a long list of feature requests which were unaddressed in the most recent development sprint (March through August). DEIMS' presence at professional meetings and conferences are vital for community adoption.

Just like the White House and thousands of others, we contribute back to the Drupal community: some of the extensions sponsored by the NSF-funded LTER DEIMS are already in use by other projects. The co-author list of this article would run in the hundreds if we were to include the work of developers that have contributed to DEIMS either by contributing to the Drupal core, or any of the 80+ Drupal contributed modules that DEIMS leverages. We invite all LTER sites to actively participate in DEIMS, as well as any person or group interested in being a part of a common solution to all information management. The group effort is precisely the main strength of DEIMS: a unified approach and solution to information management for sites, stations, and research projects.

Citations:

Apache Foundation. "Apache Solr". Accessed December 2013 http://lucene.apache.org/solr

Dries, B. "The Drupal Content Management System". 2001. Accessed Dec.2013 at http://drupal.org

Gries, C.; San Gil, I.; Vanderbilt, K.; and Garritt, H. 2010. Drupal developments at the LTER Network. Databits, Spring 2010. http://databits.lternet.edu/spring-2010/drupal-developments-lter-network.

Porter, J. LTER Controlled Vocabulary Working Group planned. Databits, Spring 2013. http://databits.lternet.edu/spring-2013/lter-controlled-vocabulary-workshop-planned

San Gil, I. 2011. The Drupal Ecological Information Management System (DEIMS): Recent Progress and Upcoming Challenges for a Grassroots Project. Databits, Fall 2011. http://databits.lternet.edu/fall-2011/drupal-ecological-information-management-system-deims-recent-progress-and-upcoming-challen.

San Gil, I. 2013. The New Drupal Ecological Information Management System (DEIMS). Databits, Spring 2013. http://databits.lternet.edu/spring-2013/new-drupal-ecological-information-management-system.

San Gil, I.; White, M.; Melendez-Colom, E.; and Vanderbilt. K. 2010. Case Studies of Ecological Integrative Information Systems: The Luquillo and Sevilleta Information Management Systems. Communications in Computer and Information Science 108:18-35. DOI: 10.1007/978-3-642-16552-8_3. http://www.springerlink.com/content/j183x10588574846/.

Shao, K.T., Lai, K.C, Lin, Y. C., Chen, L. S., Li, H.Y., Hsu, C.H., Lee, H., Hsu H. W. and Mai G. S. "Experience and Strategy of Biodiversity Data Integration in Taiwan" Data Science Journal 12 (2013): 27

Tunkelang, D. 2009. Faceted Search, Synthesis Lectures on Information Concepts, Retrieval, and Services. Wikepedia Vol. 1, No. 1, Pages 1-80 (doi: 10.2200/S00190ED1V01Y200904ICR005).

Commentary

Efficient Data Curation with the EML Congruence Checker

edit

Margaret O'Brien (SBC)

First envisioned at the ASM in 2009 and in production for a year now, the EML congruence checker is paying off. It’s something we can point to proudly as a Network: a tool that make data curation more efficient and helps ensure higher quality at all sites.

A goal at SBC was to have all data packages uploaded to PASTA by the end of 2013, so I’ve spent much of December immersed in data, EML, and the data package congruence checker. Today, SBC has 156 data packages with 286 entities; 80% of the entities are EML “dataTable”. Many are legacy packages that were originally built with Morpho (before I started at SBC), and recently imported into Metabase (by me). I anticipated that getting through them all would be a grueling, difficult slog. The most intense work took only about 5 days, and everything has made it through.

Having helped write the requirements, I knew that the congruence checker was up to the task. But what I didn't know until I immersed myself in the reports was how thoroughly and carefully Duane Costa had answered our requests for certain features. Duane has written the most submitter-friendly software that I could imagine.

Sometimes finding errors is like peeling an onion: you find one error, fix it, and then another one is exposed. So the IMC asked Duane for a feature: "during evaluate, don’t stop on the first error. Tell the user as much as possible about the data package before it stops". Duane made this work. For example, when you run in "evaluate" mode, you might learn that you have a) invalid EML, b) no dataset abstract or keywords, and c) the EML metadata lists 11 dataTable attributes, but 3 of the rows have only 10 fields. In typical checking schemes, the invalid EML would have stopped the process. After I’d fixed that problem, I’d then see the first short-row error. Then I’d have to process the package twice more to see the other two errors. But now with the the congruence checker, you don’t have to peel the onion. Duane’s code does it for you, and has given us a terrific advantage.

Yes, some of the error messages are still a little cryptic – but it’s the IMC’s job to come up with the most useful language for those, not Duane’s. And now that we have a framework, we could start the exciting part: we could validate ranges of measurements, or provide descriptive summary statistics of the data values themselves -- something a data consumer will appreciate.

The checker is helping us reduce the cost of curating dataset updates. We can now hand off the task of evaluating datasets prior to submission to a part-time assistant, and have the high-level of assurance that if a dataset passes "the checker", that it is known to be structurally correct. This code was written as part of the EML suite of utilities, which means it can be used by anyone needing to proofread EML datasets, not just LTER.

Network Information Management Products Roll Out in 2013

edit

John Chamblee (CWT) and Philip Tarrant (CAP)

Chamblee and Tarrant are the Co-Chairs of LTER Information Management Committee

LTER Information Managers have had a watershed year for the release of community-developed products for Network use. With the ramp-up in production-oriented working groups and workshops funded under the American Recovery and Reinvestment Act (ARRA) Operational Plan, the momentum began building for LTER-generated environmental informatics tools. Early products included the LTER Controlled Vocabulary (http://vocab.lternet.edu/vocab/vocab/index.php) and Unit Registry (http://unit.lternet.edu/unitregistry/) tools, as well as a large number of plans and prototypes. This year began with the release of PASTA. Other new tools have followed in a steady stream ever since.

As 2014 approaches and we consider the future of the LTER Network Office, centralized information management approaches, and broader questions concerning LTER Network structure, we should look around and see what we as a community can accomplish when we make the commitment to to build common tools to solve common problems. This article provides a list that includes many of our new network products. While incomplete, the list nevertheless highlights what we have done and serves to inform what we can do when we work together to develop creative approaches to emerging challenges.

PASTA Data Packages: Since PASTA's release by LNO in January, LTER Information Managers have contributed over 3,300 data packages. Each quarter the number of packages contributed increases and, overall, the LTER Network looks to be on-target for migrating the vast majority of LTER data to PASTA by year's end.

DEIMS: This summer, the DEIMS team presented the Drupal Ecological Information Management System as a finished product, fully capable of both producing data sets for use in PASTA and ingesting EML-described data for publication in local systems. DEIMS is being used by six sites and is now available for use, not only within the Network, but through Drupal's main development site for anyone with a need to manage ecological research data. As this article went to press, the Drupal site reported 269 downloads and seven sites using the software, meaning that at least one site beyond the LTER has adopted the freely available IM tool.

GeoNIS: What began as a broad effort to deal with the role of spatial information in the archiving of LTER data has resulted in a PASTA-based REST-ful web service capable of serving spatial data both over the web and using off-the-shelf GIS analysis software. The GeoNIS is now fully available as a service for those wishing to manage and distribute GIS data. Its successful deployment is an excellent example of how the PASTA Application Programming Interface (API) can be used to build value-added products. As more GIS data makes its way into PASTA, the value of the GeoNIS will continue to increase.

The GCE Data Toolbox for Matlab: The GCE Data Toolbox for Matlab has been in production for well over a decade, but following a 2012 training and development workshop, its use has increased dramatically. In addition, the documentation, usability, and community support for this metadata-driven data analysis engine reflect its growth. Among the accomplishments surrounding the Data Toolbox are a suite of tools integrating Data Turbine with the CUAHSI Hydrologic Information System. Wade Sheldon has a good article in this issue of Databits highlighting the uses to which the Toolbox is being put.

The Network Information System Reporting Tool: Built through a close collaboration between LNO and the IMC, the NIS Reporting system represents a key milestone not only because it contains up-to-the-minute information for anyone interested in LTER contributions to our data repository, but also because it represents another successful effort to build a value-added product on top of PASTA's API.

PASTA DOIs and Data Citation: James Brunt recently informed us that he ran across a peer-reviewed publication that included a reference to a PASTA DOI. In "Ecology and Evolution Affect Network Structure in an Intimate Marine Mutualism", published in Volume 182, No. 2 of *The American Naturalist*, (pp. E58-E72), Andrew R. Thompson, Thomas C. Adam, Kristin M. Hultgren, and Christine E. Thacker note that their data has been deposited into the MCR data catalog and provide the PASTA DOI.

As we think about and debate the future of the LTER Network structure, it is important to reflect on what we have already done. One question on the minds of many is how we can capitalize on economies of scale by centralizing information management practices as appropriate. A review of the tools presented here suggests that many of the components needed for such efforts exist and that, rather than focusing on the technical details of tool construction, we may be well served to step back and consider the management and structural changes that will be needed to federate and organize these tools and ourselves into a system that is robust and flexible enough to support us in our varied institutional and ecological settings.

Understanding the True Cost of LTER Information Management

edit

Philip Tarrant (CAP)

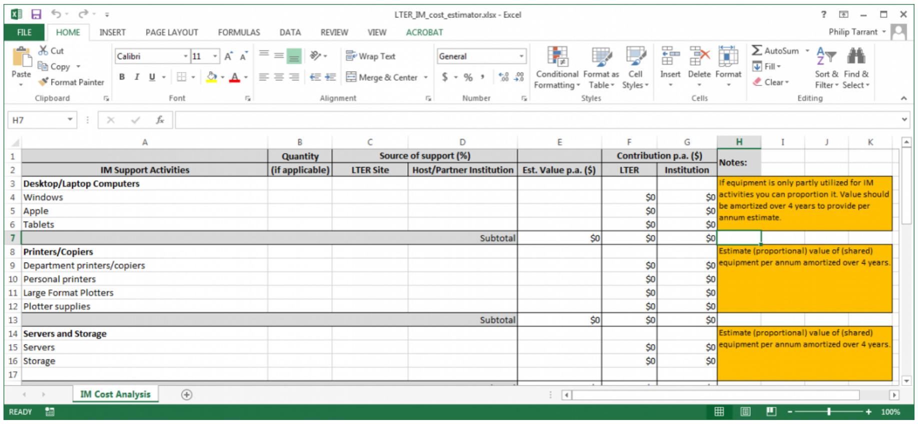
Background

Information management (IM), as a function within the LTER network, is undergoing significant change. Early implementation of the PASTA framework within the Network Information System (NIS), has required site Information Managers to reevaluate their data catalogs in light of the quality standards necessary for successful data ingestion. This activity, combined with centralized information management workshops at the 2012 All Scientists' Meeting, and possible future changes in LTER network operations, mean that different information management models have become a regular discussion topic. However, in order for us to have an informed debate on this subject, we should take the time to try and understand the true cost of information management as it exists today.

The current models for site information management vary significantly dependent on how the LTER site is supported by its host institution. Some sites operate as stand-alone research stations. Other sites are heavily integrated within their host institution. Yet others are effectively multi-agency collaborations that serve a diverse set of "masters". This diversity of organizations is likely to lead to a variety of service delivery models and it is unlikely that a "one size fits all" formula would accurately estimate costs for the entire network. Therefore, it makes sense for each site to assess its own IM costs within the context of how the site is organized.

The Central Arizona-Phoenix LTER site is hosted by the Global Institute of Sustainability at Arizona State University. The Institute's Informatics and Technology team provides the information management services for this site as part of a larger portfolio of technology services. As CAP LTER is by far the team's largest customer, I felt it was important to understand how the support provided by the team related to the contribution made by the project. Consequently, I set out to estimate the cost of support operations in terms of personnel, hardware and software.

Information management cost estimator



	A	B	C	D	E	F	G	H
1		Quantity	Source of support (%)			Contribution p.a. ($)		Notes:
2	IM Support Activities	(if applicable)	LTER Site	Host/Partner Institution	Est. Value p.a. ($)	LTER	Institution	
3	**Desktop/Laptop Computers**							If equipment is only partly utilized for IM activities you can proportion it. Value should be amortized over 4 years to provide per annum estimate.
4	Windows					$0	$0	
5	Apple					$0	$0	
6	Tablets					$0	$0	
7				Subtotal	$0	$0	$0	
8	**Printers/Copiers**							Estimate (proportional) value of (shared) equipment per annum amortized over 4 years.
9	Department printers/copiers					$0	$0	
10	Personal printers					$0	$0	
11	Large Format Plotters					$0	$0	
12	Plotter supplies					$0	$0	
13				Subtotal	$0	$0	$0	
14	**Servers and Storage**							Estimate (proportional) value of (shared) equipment per annum amortized over 4 years.
15	Servers					$0	$0	
16	Storage					$0	$0	
17								

To make this calculation I created a worksheet to estimate costs and then proportion those costs between the project and the Institute. This worksheet is an MS Excel spreadsheet with embedded formulas that contains sub-sections for:

- Desktop/laptop computers
- Printers/copiers
- Servers and storage
- Specialist software products
- Education and travel
- Web support
- IM related personnel
- Supplementary personnel

While these sections were all relevant for CAP LTER, they would not all be completed at sites where the Information Manager wears several hats (e.g. database management and web site maintenance). The worksheet also requires a number of assumptions to be made. Although some of these assumptions are vague (which increases the error range), it still allows for a reasonable estimate to be produced.

The assumptions used for the CAP LTER assessment were:

Hardware (PCs, servers, printers, etc.)	If the equipment is only partly utilized for LTER activities it should be apportioned. Value should be amortized over 4 years.
Software	Per seat license costs are difficult to estimate. License costs vary with volume and many suppliers will only provide quotes on requests. Value should be considered as that the LTER site could negotiate as an individual entity rather than the volume discount available to the institution.
IM personnel	Salary contribution should be the effort required to support LTER IM activities only. This number will be (% of time spent on IM activities) x (salary + benefits). The dollar amount of support that is provided by the LTER site and the institution can then be apportioned.

Conclusions

After completing the worksheet I determined that the CAP LTER project contributes approximately 22% towards the cost of services and support received. While this contribution may appear low, it can actually be considered as good business for the Institute. This is primarily because, as our largest customer, the LTER project can be used as a development platform for solving Institute business problems, i.e. providing data management solutions for CAP LTER allows me to solve the problem for my other customers. To date, this has proven to be a very successful formula.

Interestingly, CWT also completed a cost assessment using the same worksheet. The Information Manager concluded that the project contributed ~89% of the cost of IM support; a very different result to CAP. However, it should be noted that information management does not exist in a vacuum and LTER projects receive their institutional support in a variety of ways. Differences in the proportion of IM support received from a host institution should always be viewed in the broader context of the complete LTER site operation.

This symbiotic relationship between LTER sites and their host institutions creates an interesting dilemma with respect to information management, as we go forward. Any dialog we have with respect to possible changes in the LTER information management model should seek to understand the net dollar benefit received by the LTER network through these close institutional associations. In addition to this benefit we should ensure we appreciate the actual cost of alternative service delivery methods. If we approach this topic with due diligence we are less likely to make choices that ultimately prove to be more expensive than anticipated.

News Bits

GeoNIS Project Update

edit

Theresa Valentine (AND)

The GeoNIS working group received funding from the LTER Network Office (LNO) to automate a workflow to bring PASTA data with a spatial component (GIS and imagery data) into a centralized GIS formatted database and provide map and image web services for all LTER sites. The spatial data are quality checked against EML (Ecological Metadata Language), and error reports are generated for submitters.

The contractor created a user interface where sites can view their spatial data, link to map and image web services, and view a report on their data (check for errors). The programs are developed as tools within the esri toolbox structure (programed in Python), and are located on a server in the LNO office. The application is programmed with the ArcGIS Javascript API, and is available at the following link: http://maps3.lternet.edu/geonis/.

Specifications for the projects were developed at a meeting with the contractor, GeoNIS working group, and LNO staff in January of 2013. Additional on-line meetings were held throughout the contract period, and a demonstration of the capabilities of the system was presented to the Information Managers Committee and during a working group breakout session at their 2013 annual meeting in Fairbanks, Alaska.

Following discussions with LNO staff at the LTER Information Managers Committee meeting in Fairbanks, a value-added product was discussed to provide a geoprocessing service that PASTA could initiate for a data check, prior to data being placed in the production PASTA portal. The current GeoNIS tools work in production and testing mode. It would be possible to have sites submit data to the testing mode before submitting their data to PASTA. The contractors did not have time to complete this task on this contract. They did outline a process for the task, and estimated that it would take about 80 hours to complete the task at a cost of $3,200.

The GeoNIS working group has committed to monitoring the project, and will be moving it to production mode (where the workflow runs on a schedule, searching PASTA for new spatial data, and running the workflow and updating the web services).

The GeoNIS working group would like to acknowledge the support it received from LNO to fund this development. The project has helped sites document their spatial data with EML and to incorporate these data into PASTA, and provides mapping services to the sites for their spatial data. In addition, several LTER sites provided sample datasets for testing, and worked to improve the workflow and best practices for documenting spatial data.

Members of the GeoNIS working group are: Adam Skibbie (KNZ), Theresa Valentine (AND), Jamie Hollingsworth (BNZ), and Aaron Stephenson (NTL). Contractors were Ron Beloin and Jack Peterson.

Good Reads

Review: Common Errors in Ecological Data Sharing

edit

Hope Humphries (NWT)

Kervin, Karina E., William K. Michener, and Robert B. Cook. 2013. Common errors in ecological data sharing. Journal of eScience Librarianship 2(2):Article 1. http://dx.doi.org/10.7191/jeslib.2013.1024.

This study identifies common errors in data organization and metadata completeness that were discovered by reviewers of data papers published in the Ecological Society of America's (ESA) Ecology Archives. ESA's Ecology publishes the abstract of a data paper; Ecological Archives contains the data sets themselves and accompanying metadata, allowing for long-term access to data sets, which can be updated. An average of about 20 errors was identified per data paper, although many of these were simple editing errors. The authors grouped the errors according to the Data Life Cycle elements described by Michener and Jones (2012). Over 90% of papers had errors in the Collection and Organization category (i.e., collection methods, site/time descriptions, inclusion of all relevant variables) and the Description category (i.e., ascribing metadata to the data). The pervasiveness and number of errors in the data sets analyzed is perhaps surprising considering that they were specifically submitted for publishing in a data archive, and therefore one might expect that extra attention had been paid to their completeness prior to submission. However, the careful scrutiny given these data sets by reviewers no doubt was a factor in unearthing problems that might otherwise have gone unrecognized.

Information managers are all too aware of the existence of missing information and errors in data and metadata, but this study's results could be used to raise awareness in scientists and students about the kinds of problems they should be on the lookout for. As a best practice, the authors emphasize the importance of recording all details about the study context, data collection, quality control, and analysis throughout the course of a research project rather than attempting to reconstruct them after the fact. Saving data in a non-proprietary format, such as ASCII text, is noted as important. Organizations that offer data management training and educational tools are also identified.

Reference:

Michener, William K., and Matthew B. Jones. 2012. Ecoinformatics: Supporting ecology as a data-intensive science. *Trends in Ecology & Evolution* 27(2):85-93. http://dx.doi.org/10.1016/j.tree.2011.11.016.

Review: Big Data: A Revolution That Will Transform How We Live, Work and Think

edit

John Porter (VCR)

"Big Data" poses major challenges in perspective for information managers and scientists alike. The book "Big Data: A revolution that will transform how we live, work and think" by Viktor Mayer-Shönberger and Kenneth Cukier does a good job of laying out opportunities and challenges afforded by "Big Data." There are some aspects of their advocated approach, specifically the emphasis on correlation over causation, that represent major challenges to our science. To quote the book: "The ideal of causal mechanisms is a self-congratulatory illusion; big data overturns this." Nonetheless, the power of "Big Data" is, and will, influence how we manage and analyze data within ecology.

The book is written for the non-technical reader and is replete with a wide array of real-world examples of the application of "Big Data." The chapter headings of the book "Now," "More," "Messy," "Correlation," "Datafication," "Value," "Implications," "Risks," "Control," and "Next" say a lot about Big Data by themselves, emphasizing the velocity and magnitude of data collection, how when you have lots of data, you can live with less accurate or precise data, the power of prediction, and the opportunities and challenges of Big Data.

As elsewhere, the definition of "Big Data" is a sliding scale, being defined more by the need to apply non-traditional analytical approaches than the size of the data itself. A strength of the book is the wide variety of examples used. Many of them come from the world of massive data derived from Internet search engines, or web crawlers (e.g., Google's ability to predict flu outbreaks based on search terms used), and social media. However, there are also examples that are less anthropocentric, such as the application of many low-cost, low-accuracy sensors in place of a few high-cost, high-accuracy sensors.

A thesis of the book is that many of the rules that applied in the past no longer apply today. For example, when you can collect essentially all the data from a population, sampling becomes unnecessary. Similarly, when you have really massive amounts of data, irregularities in merely large amounts of data become inconsequential.

As a scientific reader worshipping at the altar of accuracy and precision, there was much in the book that seemed radical or even subversive. However, although I don't think that "Big Data" approaches will entirely supplant traditional sampling approaches, they can augment and enhance our ability to address ecological problems, allowing us to attack problems that are not soluble using traditional scientific approaches.

Theme by Dr. Radut.