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

Fall 2010

Welcome to the Fall 2010 issue of Databits wherein 18 authors contribute from 11 sites and 2 institutes in 3 countries. Seven relevant and insightful feature articles relate to growth and refinement of systems. All three editorials share the theme of continuous change. And both the suggested readings reflect on past adaptations. Even the two tools recommended prove their worth in how they handle revision.

Keeping pace with evolving standards in information management may seem like “it takes all the running you can do, to keep in the same place." – Lewis Carroll. This ongoing challenge is captured in the Red Queen Principle, proposed by the late professor emeritus L. van Valen in 1973: “for an evolutionary system, continuing development is needed just in order to maintain its fitness relative to the systems it is co-evolving with.” The LTER Network has remained relevant through 30 years of change and was just last May recognized for its contributions by the American Institute of Biological Sciences. We continue to raise expectations for our practice. Within the large scope of information management we may feel small and progress may appear slow. Databits captures a twice-annual snapshot of information management co-evolving with Network needs. Now that all years’ issues are gathered within one site, the progress of LTER IM practice may be viewed like time-lapse photography. 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.

----- Co-editors: M. Gastil-Buhl (MCR), Philip Tarrant (CAP)

Featured Articles

Commentary

News Bits

Good Tools And Programs

Good Reads

FAQ

Calendar

Featured Articles

Coweeta Supports Citizen Science Initiative; Collaboratively Redevelops and Publishes Long-Term Biomonitoring Database

edit

John F. Chamblee (CWT)

Since 1990, The Little Tennessee River Biomonitoring Program, housed at the Little Tennessee Watershed Association (LTWA) has provided a record of stream health for the Upper Little Tennessee River and its tributaries. With the help of thousands of volunteers, Biomonitoring Program Director Dr. William O. McLarney, has been sampling fish to develop species inventories at stream sites across the Upper Little Tennessee River basin. The resultant database is one of the largest of its kind and includes observations for hundreds of thousands of individual fish representing all species known from the watershed. Most samples include a record of fish counts (by species) and a stream health ranking also known as an Index of Biotic Integrity (IBI) score. Each year, the Biomonitoring Program uses these data to issue “State of the Stream” reports that serve as a means to provide community leaders with feedback on water quality and general ecosystem health.

In 2008, Dr. McLarney approached the Coweeta LTER with a request for assistance in making the data public in an easily accessible format. Prior to his request program data were stored in a single, loosely structured excel spreadsheet (Figure 1) and directly accessible only to Dr. McLarney. In response to Dr. McLarney’s request, Coweeta offered to re-organize the data into a relational database framework and develop a simple web application for data download. In exchange, the LTWA provided assistance with quality assurance and quality control on the data and the right to curate the data in our data catalog and serve the data to the public through our web servers.

Over a two-year period, we worked collaboratively with Dr. McLarney, LTWA Executive Director Jenny Sanders, and North Carolina Department of the Environment and Natural Resources Watershed Planner Andrea Leslie to migrate data from McLarney’s semi-structured data management system to an integrated RDMS. Coweeta IM staff developed the work plan and a series of interim data models that would allow LTWA staff to crosswalk the data through successively more complex data structures using the Excel Spreadsheet software with which they were familiar. Once this work was complete, I imported the data into a new data model, developed in Microsoft Access (Figure 2). At the same time, Andrea Leslie worked with Dr. McLarney to provide spatial locations for all of his 368 sampling locations.

Once the data were in the new database, staff from the LTWA and CWT IM worked with Anya Hinkle, Assistant Director of the Highlands Biological Station, to oversee student interns as both staffers and students performed a QA/QC check on all 8,000 species entries, over 500 IBI ranks and scores, and all site locations and site descriptions. These checks were specifically geared toward ferreting out transcription errors that developed from the transfer of data from a semi-structured format to a relational one. Finally, the data were transferred via an online MySQL database so that CWT IM staff could develop a web interface.

Over and above the fact that we have established long-term community ties centered on an important citizen science initiative, this project has achieved three substantive outcomes thus far. The students from Highlands Biological Station produced a report that includes a preliminary comparison of the data against Coweeta's forthcoming 2006 land cover classification. That report is available here. More importantly, however, has been the collaboration's contribution to citizen-science-based decision-making by the North Carolina government. By providing a quality assured copy of the data to the North Carolina Natural Heritage Program (NCNHP), Coweeta and the LTWA were able to jointly provide the NCNHP with the data they needed to significantly expand the spatial footprint of the aquatic species diversity heritage zone for the Upper Little Tennessee Watershed by adding several previously unlisted tributaries. Finally, with the launch of a simple website to showcase the data and facilitate download, we have achieved Dr. McLarney’s goal of making the data publicly available to everyone and have added an important long-term data set to the LTER catalog.

Researchers at both the LTWA and Coweeta view this effort as merely the beginning of a long-term process that will enhance awareness of biodiversity locally while increasingly involving students and the general public in the setting and implementation of conservation priorities. The database will be updated on an annual basis using a canonical copy of the Microsoft Access application (Figure 3) that will be provided by Coweeta. This platform has been selected in recognition of the LTWA’s location in a rural area with occasionally unreliable Internet service. Annual versions of the database will be maintained in Coweeta’s new data catalog (to be released in early 2010) and the on-line database will continue to be updated with a series of ordered SQL operations and an ODBC connection. Over the long run, Coweeta will support the LTWA as they seek funding to enhance the website and we will continue to curate the database, update it, and maintain the website on the LTWA’s behalf.

Implementing ProjectDB at the Georgia Coastal Ecosystems LTER

edit

Wade Sheldon (GCE), John Carpenter (GCE)

Background

Two LTER workshops were convened in 2008-2009 to plan and develop ProjectDB, a cross-site research project description language and database. The first workshop brought together a diverse group of LTER Information Managers to define the scope and use cases for the database (Walsh and Downing, 2008; O'Brien, 2009). A second workshop was convened in April 2009, where a smaller group developed an XML schema for storing the targeted project information (lter-project-2.1.0, based on eml-project) and prototyped XQuery-based web services using eXist, a native XML database system (Gries et al., 2009; Sheldon, 2009). The ProjectDB effort was very effective and serves as a model for collaborative software design in LTER; however, design is not the end-point in the software development process. This article describes taking ProjectDB to the next level at the Georgia Coastal Ecosystems LTER site (GCE), by putting the schema and database into production and integrating it with the rest of our information system.

Planning for Implementation

The first decision we had to make was how the concept of a "research project" should be applied within our site organizational scheme. For example, should every funded (or published) activity of every project investigator (P.I.) and graduate student be considered a separate research project, or should we only document major focus areas of the larger GCE research program as projects? After discussing the possibilities with the GCE leadership, we settled on a simple approach: any research activity that is specifically named in an NSF proposal or supplement and that is addressing a stated research question from a GCE proposal will be considered a research project. We felt that tying projects to questions and elements of proposals would best serve our need to track progress on all funded activities, and would provide NSF and reviewers the clearest alignment between proposed activities and research products. This approach also mirrors how we have traditionally assembled annual NSF reports, thereby simplifying constructing project descriptions retrospectively (see below) and allowing us to use ProjectDB to organize report contents moving forward.

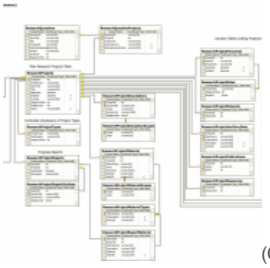
Next we had to decide what specific information to include in project descriptions. The ProjectDB schema, like its parent EML, is highly flexible and can store a wide variety of content at varying levels of detail (O’Brien et al., 2009). For our initial implementation we chose to focus on descriptive abstract text and links to GCE personnel, research questions, publications and datasets as the highest priority content. Overall geographic and temporal coverage and keywords were also considered critical for search purposes. We additionally wanted to include supporting imagery (e.g. graphs, maps) and links to information outside the GCE information system to allow investigators to showcase their study designs and results informally but effectively. Management information (e.g. project review and approvals), fine-grained geographic information and cross-links to related research projects were deemed less important for this initial implementation, though, and were deferred to a later release.

XML vs. Relational Content Models

Our next decision was how to implement the back-end content model for the database. The ProjectDB design focused solely on XML technology (XML schema, XQuery, XSLT, eXist), based on broad computer platform and language support and to leverage prior work on the EML schema and pre-built functionality in the eXist XML database (e.g. REST web services and document management). However, like many LTER sites we manage all GCE metadata and research information in a relational database management system (RDBMS). While there are distinct advantages to both document-centric (XML) and record-centric

(RDBMS) approaches to content management, bridging between these two models can be complex and has ramifications for how dynamic web content is generated and searched. For example, XML documents are typically displayed and manipulated using XSLT (or XQuery), whereas server-side web applications or client-side AJAX applications are more often used to render and search relational database contents on the web.

After much consideration, we chose to extend our existing RDBMS (SQL Server) to manage research project description information. This allowed us to link projects to existing database content, avoid record duplication, take advantage of referential integrity, and simplify development of data entry forms (see below). Following the philosophy of the lter-project schema, we developed a generalized SQL model for storing an unlimited amount of descriptive text at different scopes, as well as unlimited reports and associated material. Controlled vocabularies for document scope and material type, as well as explicit “DisplayOrder” fields for text sections, allow content to be nested appropriately in the XML document and tagged using attributes to support differential HTML generation and styling using XSLT (e.g. image tags versus anchor tags). Junction tables are then used to define many-to-many relationships between research projects and GCE questions, personnel, data sets, publications, archived files, geographic locations, taxonomic database records and external web resources. The entity-relationship diagram of the database is shown in figure 1 and also available on the GCE web site (http://gce-lter.marsci.uga.edu/public/app/resource_details.asp?id=376).



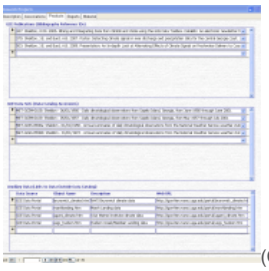
(Click on image to enlarge)

Figure 1. Entity-relationship diagram of the research project schema in the GCE_Submissions database.

We chose to use XML technology, though, for developing the remainder of our system to leverage and contribute to the broader ProjectDB effort. To accomplish this SQL-XML hybridization, we developed database code and server-side ASP web applications to dynamically generate lter-project XML from SQL Server using the same approach and code library used to generate dataset EML. In essence, we designed SQL views to generate XML fragments dynamically for each major element of the document, employing conditional logic (or subqueries and unions) where necessary to support complex-type elements. These views are then queried in an ASP script, and the fragments are assembled in document order and streamed to the client to create an XML document. SQL functions are used in views to escape illegal characters that may appear in paragraph contents (e.g. <,>,&), and an IIS code-page directive is used to encode text as UTF-8 prior to streaming for XML parser compatibility. Many other approaches can be used to achieve similar functionality, including native XML support in SQL Server and other major RDBMS software, but this approach has proven effective and easy to manage through years of use and refinement at GCE.

Managing Content

As mentioned above, a significant advantage of storing content in a relational database is ease of developing form-based data entry interfaces. Although XML-based form generators are emerging (e.g. XForms), relational database tools have matured over decades and many options are available for rapid development of client-server and web form interfaces. We chose to use Microsoft Access to develop an initial set of client-server forms, in order to take advantage of its rich set of data-bound controls and excellent SQL Server support. Based on past experience, we anticipated that only GCE IM staff and project lead P.I.s would initially enter and update project descriptions, so the platform and software limitations imposed by Access were an acceptable trade-off for its rapid application development capabilities. In the future we plan to deploy web-based forms for end-user updates, as we do for bibliographic citations, document archiving and other user contributions on our web site. Screen shots of representative tabs on the Access-based forms are shown in figure 2.

(Click on images to enlarge)

Figure 2. Microsoft Access user interface forms used to enter and update information in the SQL Server database.

We initially populated the database using content from annual NSF progress reports for the first 9 years of the GCE project. Project titles, associated personnel, descriptive text and figures were extracted from Microsoft Word and Adobe Acrobat files and pasted into Access form fields to create project records. Time spans and geographic coverage were derived from IM knowledge, history of report entries (i.e. which years the text appeared), and proposal details. Input from lead project P.I.s and other investigators was used to associate projects with research questions, data sets and publications. The initial population of the database took about two weeks for 75 project descriptions. After filling in as much information as possible, a request was sent out to all GCE participants to review their project entries and submit edits and updates to the IM office. Although tedious, this exercise was successful and resulted in a database that was immediately useful and represented the complete history of GCE research.

Tying Everything Together

In order to support the XQuery-based web services and search form developed by the ProjectDB working group, we deployed an eXist database instance within our information system. (Note that the LTER Network Office does not currently operate a production version of eXist, so this was a necessary step). We hardened the database for production use following best practices described in the documentation (e.g. disabling public write on collections, restricting access to administrative interfaces and reverse-proxying /exist/rest and /exist/xmlrpc to Apache virtual directories). We then developed a Python script to retrieve ASP-generated XML documents and upload them to eXist using HTTP PUT in order to synchronize content between databases on a regular basis. For web display of project descriptions, we augmented the XSLT stylesheets developed for ProjectDB to import a second XSLT that creates the GCE web scaffolding and imports navigation menus via AJAX. In addition, a second "plain" stylesheet without web scaffolding and navigation was developed for printing research project descriptions or loading contents into word processing programs for content re-use in reports and proposals and to simplify end-user content revision.

The GCE implementation of the ProjectDB search form and display pages are online at http://gce-lter2.marsci.uga.edu/exist/rest/db/projects/util/xquery/getProjectsQueryForm.xql?xslUrl=http://gce-lter2.marsci.uga.edu/exist/rest/db/projects/util/xslt/gceQueryForm.xsl. In addition, conventional ASP web application pages are available for viewing a list of all ongoing and completed projects (http://gce-lter.marsci.uga.edu/public/research/projects.asp) and listing research projects addressing each GCE question (e.g. http://gce-lter.marsci.uga.edu/public/research/gce2_q1.asp). For the latter two examples, summary information is queried from the SQL database for display on these pages using ASP, but the ultimate document links return XML with an XSLT declaration to produce identical output to the XQuery results serialized by eXist.

Conclusion

We initially developed and deployed the GCE ProjectDB during preparations for our October 2009 site review, and it proved invaluable for that purpose. Coupled with new dynamic web pages describing GCE research questions (see http://gce-lter.marsci.uga.edu/public/research/research.htm), both GCE investigators and reviewers could easily determine how specific research elements were contributing to our overall mission. Dynamic hyperlinks between research projects and personnel pages, data set metadata, and other database content also provide more ways to navigate to related content on the GCE web site. As a cautionary note, though, this database provided reviewers with a birds-eye view of our productivity (or lack thereof) on specific project elements. However, this increased transparency and project traceability will be crucial to identify holes and weak areas that need attention as we prepare for our renewal proposal.

In the year since the review we have found ProjectDB to be a very effective tool in our information system. This database provides an essential link between research activities, people, and products just as envisioned during the LTER planning workshop. Although keeping this database current will require ongoing effort as new projects, data sets and publications are added, the ability to leverage this database to generate text for annual reports and for reviewing working group progress more than offsets this work. We have uploaded our site-specific modifications to XQueries and XSLT stylesheets to the ProjectDB development instance of eXist at LNO, and look forward to continued collaboration on developing this cross-site resource.

References

Gries, C., Bohm, S., O'Brien, M., Porter, J., Sheldon, W. and Walsh, J. 2009. Project Database for LTER Sites. LTER Information Management web site (http://intranet.lternet.edu/im/project/LTERProjectDatabase).

O'Brien, M., Bohm, S., Porter, J. Gries, C., Sheldon, W. and Walsh, J. 2009. The LTERProject XML Schema. LTER Information Management web site (http://intranet.lternet.edu/im/project/LTERProjectDatabase/documentation...).

O'Brien, M. 2009. Lessons learned from the projectDB workshops. LTER Databits – Information Management Newsletter of the Long Term Ecological Research Network, Spring 2009 (http://databits.lternet.edu/spring-2009/lessons-learned-projectdb-workshops).

Sheldon, W. 2009. Getting Started with eXist and XQuery. LTER Databits – Information Management Newsletter of the Long Term Ecological Research Network, Spring 2009 (http://databits.lternet.edu/spring-2009/getting-started-exist-and-xquery).

Walsh, J. and Downing, J. 2008. PROJECTDB – Planning and Development of a Collaborative Programming Effort. LTER Databits – Information Management Newsletter of the Long Term Ecological Research Network, Fall 2008 (http://databits.lternet.edu/fall-2008/projectdb-).

IMC Governance Working Group: Developing a Terms of Reference

edit

Karen Baker (PAL, CCE), Nicole Kaplan (SGS), and Eda Melendez-Colom (LUQ)

The concept of IMC governance: its origin and goals

At the 2008 Information Management Committee (IMC) annual meeting, the need to define ourselves more formally as an LTER IM Committee emerged. Recognition had been growing in prior years that new arrangements were required to manage new demands on the committee activities and resources. Requests for participation in LTER Network IM activities have been increasing as well as the expectation for partnering outside the LTER Network. The committee's growth in size and responsibilities made apparent the value in better defining our work methods and criteria for decision-making as well as in broadening our traditional informal style to include more formal governance. The Governance Working Group (GWG) was formed at the 2008 IMC annual meeting with the initial goal to explore, document and learn from the ways IMC members have conducted their governance. The desire to preserve valued aspects of the IMC was strong despite plans for changing practices to meet changing circumstances. The GWG effort involved identifying elements of governance, by reviewing governance structures, practices, and decision-making. The GWG was charged with recommending governance practices for the IMC, recognizing that such an effort could be informed by revisiting past and existing practices.

At the 2008 meeting, there followed an intensive 48 hours of developing interview questions and conducting interviews of four past IMC chairs (Susan Stafford, Barbara Benson, Nicole Kaplan, and Corinna Gies) together with an Executive Board Representative who was an early committee participant (John Porter) – all of whom happened to be at the meeting. Subsequently, a GWG meeting was arranged synergistically with the LTER SocioEcological Science (SES) working group in Puerto Rico in December 2008. In developing the goals and milestones of the GWG and beginning to identify governance principles and established past practices through analysis of interview transcripts, we were able to provide input to the SES breakout groups as well as to refine the GWG charge and elaborate on the governance structure and roles within IMC. Two IMC virtual 'watercooler' videoteleconferences (VTC) in March and August of 2009 kept the IMC informed and engaged with GWG activities. Our approach began to take shape, and it was clear it would address the decisions made by the IMC and the IMC steering committee, IMExec. GWG VTCs were devoted to reviewing strategies for decision-making, prioritizing projects, and coordinating of IMC activities, expanding the organizational context explicitly to reflect LTER by-laws, outlining communication with a network information system advisory committee, and responding to new mandates and expectations from National Science Foundation representatives.

Governance models and principles

Governance may be defined as a set of ideas about how direction is provided to collective human activity. We identified appropriate definitions of governance to ground our efforts using information gathered earlier by the Network during their strategic planning process. The Network governance process was reviewed at a series of LTER Governance Working Group meetings in 2005:

"*What distinguishes governance from management is that it is concerned with how the big (or strategic directions) are taken… The process of governance typically rests on a governance system or framework. The formal elements of this system (constitutions, bylaws, conventions) define how the process is supposed to function in a particular setting. But in practice, the informal decisions, accepted practices, or unwritten codes of conduct that people follow are often equally important in determining how governance works.*" (T.Plumptre; http://www.goodgovernancebd.org/link/concept_papers/Concept_papers07.htm).

A shortened version of a set of good governance principles (UN Development Program, 1997) had been summarized as follows:

- **Legitimacy and voice**: Participation and consensus orientation
- **Direction**: Strategic vision
- **Performance**: Responsiveness; effectiveness and efficiency
- **Accountability**: Accountability & transparency
- **Fairness**: Equity

These good governance principles were applied to ensure that the work of the IM GWG was grounded and our Network and IMC values preserved as we moved forward.

IMC GWG charge, historical context, and emergence of the ToR

After formation, the charge of the committee was clarified at a March 2009 VTC:

"*The IMC GWG will prepare a summary to inform IMExec and IMC about governance in terms of organizational framework and with examples of collaborative activities and mechanisms that support Network science and long-term information infrastructure. In preparing these materials, the IMC GWG will rely on information from the IMC and broader community relating to the LTER strategic planning process.*"

In August 2009 at the IMC annual meeting, presentation of a summary diagram of our community stimulated discussion. Discussion included elaboration on the types of decision-making in addition to informal approaches (i.e. formal, semi-formal, tacit, and compliance) and the types of communication available, to broaden and make visible how business was carried out by the community (e.g. surveys, white papers, requests for comment, requests for support, and requests for endorsement).

Concerned with learning from our history, a gathering of the types of activities in which we have been engaged was developed into a collection prototyped as a 'HistoryDB' module. This was recognized as a larger project to put aside until governance issues had been addressed. The incomplete collection of events served as a record or timeline of activities, a memory trace that identified actions and provided examples of decisions made over time.

By February 2010 in a presentation to IMExec, the focus moved first from the community diagram to an organizational diagram and finally to a general recognition of the need for explicit governance guidelines, e.g. by-laws or Terms of Reference (ToR). This development parallels the LTER Network development and adoption of by-laws in 2006. A first draft of elements of a ToR was presented in a June 2010 VTC that laid the foundation for full discussions of the document at the September 2010 IMC annual meeting through a series of break-out groups. The GWG itself expanded from 3 members to 7 in order to support plans to prepare a finished document for the 2011 IMC annual meeting.

Governance and the way we work given the duality of site and network levels

There are 26 sites with different research themes, habitats and scientific questions that all have a shared goal of creating and disseminating ecological data and knowledge. This presents what appears as a situated[1]-universal paradox - the goal of making choices that optimize for individual site concerns while also taking action as members of a network. This paradox requires a balance between local site and broader network goals and is a defining characteristic of the site-network model. We have a history of respecting each individual site's choices of hardware and software, design approaches and scientific tools to do their research; it's our fortune to co-exist within a network context of long-term ecological scientists that anchors our understanding and respect for multiple perspectives and configurations. It frames our dedication to taking time to listen to and learn from each other. As information managers familiar with 'making do' given limited resources and 'good enough' given the tsunami of data and data types, we are not shy about changing practices, trying new approaches, or adopting prototypes from other sites when they are adaptable to our own particular site's historical, physical and administrative reality. Governance in the form of the ToR brings into this balance, procedures for identifying and addressing how we work at the network-level and the impacts of this work on individual sites. The site-network context also creates a venue to elevate site-based or cross-site projects to contribute to the capacity of the Network, by establishing means to support the development and enactment of new approaches.

Concluding thoughts

It was an IMC 'emergent moment' when three realizations coalesced into a new working group: Nicole with the perspective of an IMC co-chair suggesting that many of the troubles arising for IMC were related to governance, Eda recognizing the value of revisiting our history by interviewing past chairs, and Karen as past member of the 2005-2006 LTER GWG contributing experience with ethnographic methods and concepts. This was an effort focused from the start on articulation work critical to functionality of an expanded IMC facing growing pressures and continuing development. Interviews and review of historical events helped in capturing the multiple perspectives and memories involved, maintaining the inclusivity characteristic of the IMC, and facilitating consideration of a joint course of action in partnership with the IM-Exec and IMC members. Sharing a review of history prompted reflection, which was critical to identifying, defining and maintaining the unique aspects as well as the intent of the IMC. It took time for an initial reticence to changing what has been working well to shift to the realization that a ToR would help maintain some defining characteristics of the IMC - characteristics such as inclusiveness, open-mindedness, engagement, collaboration, co-learning, and respect for the paradoxes inherent to the site-network model – in the light of new demands and changing circumstances. The Terms of Reference represent a change mechanism for IMC governance that expands awareness of decision-making and openness of the governance process for information management at the network level in the LTER site-network model of science.

[1]Note: The term 'situated' is used in social sciences to refer to the individuality of the local or the heterogeneity to be expected from differing circumstances such as with physical locations or human experience; situated is often used in describing technological configurations and organizational arrangements; it refers to and acknowledges local influences that create different experiences with and understandings of data, knowledge and learning.

Addressing Scaling Issues Associated with Data Access

edit

James Conners (CCE, PAL)

Overview

One of the primary features of our local data system, DataZoo, is providing queryable access to data sets. During the last year, however, we began to face two main issues with some of our larger data sets. The data accessed from DataZoo are stored in a relational database. From a web form, queries can be built and submitted for any data set. At some point, our original implementation for these queries began to strain as the amount of data stored scaled upward. Because both the query against the data base and the retrieval of results were all being performed during a single request, some queries to our larger data sets were resulting in browser timeouts. In addition, in order to provide paging capabilities for previewing the data results the queries had to be buffered through application code to be able to get information like the number of records returned, causing queries to consume large portions of server memory and to affect performance. This article describes the approach we took for both addressing these immediate issues and planning for future system architecture development.

Design

The approach taken for the redesign of our data queries was to frame it as a local web service with two primary characteristics: A level of abstraction that generalized queries to data in different storage back- ends across our system and an asynchronous client-server communication protocol to eliminate web-server execution time limits. The idea was that this web service would provide a generic interface for querying and accessing tabularly represented data, utilizing a set of defined resources that characterized a suite of activities useful for asynchronous interaction throughout the data query workflow. Below is the description of that workflow, with each step also generally defining the set of resources making up the web service as a whole:

1. Retrieve table and field information for selected data storage back-end
2. Register query
3. Intermittently check status of query, including data result information if available
4. Access data result

Although steps 1-4 comprise the entire workflow as defined for client interaction, only steps 2-3 are coordinated within the server application through the use of a message table. The message table supports communication between the registration, query, status and access processes by providing state and summary information. Each process works relatively independently, retrieving decision logic from the message table and initiated by client requests, with the exception of the query, which is started by the registration process.

Use Case

The primary use case we used during development is described as follows. The client (a web browser, typically) retrieves table and field information from the info resources and uses it to build a form for a user to construct valid queries against a particular storage back-end. The query string is constructed from the form fields and sent to the registration resource, immediately receiving a status object with a unique identifier. The client then intermittently makes requests to the status resource using this ID for updates on the current state of processing, along with other details such as a descriptive message. If the query is successful, the

next time the client requests the status of the query it will receive a message within an updated status along with additional information describing the data result, i.e. its location, size and expiration. The client can then retrieve the data using the access resource, which supports result retrieval with some optional parameters that specify range selection and specifics for field and file names.

Details

The following presents a technical description of the workflow from the service perspective, from steps 2 onward. When the service receives a request for the registration of a query (via the register resource), it creates a message with an initial state of "processing" and sends back a uniquely identified status object describing that message. It also initiates the query process, passing it the ID of the message to refer to. The query process begins execution using information stored in the message. At any time during this workflow, the client is able to access the status resource to retrieve the current state of the workflow. When the query is finished, or upon error, the process updates the message and exits. If the query was successful, a data result was written to disk as a CSV-formatted file. All information about the data result is obtained by the status resource using this file, precluding any data-result buffering with application code. Data results are then retrieved from the access resource. This resource is a layer of abstraction over the data results that provides added functionality like range specification, file name options and alternative field names. Data is always returned as CSV. Later, another web service was developed over the top of this access resource to provide re-formatting of the data result as Excel spreadsheets and other formats. Set in the configuration of this service is the time-to-live for these data results, currently set as 24 hours for our system. If either the status or access resources are accessed using a query ID after this period, the message is analyzed and appropriately set to expired and the client is notified. Below is a sequence diagram illustrating this interaction.

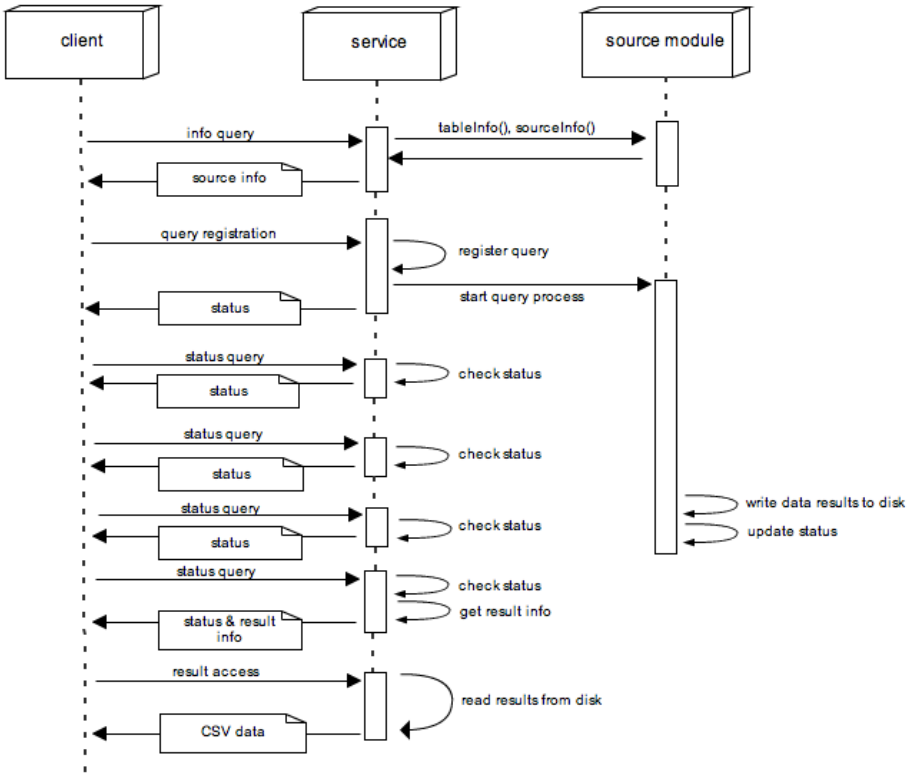


Figure 1. Sequence diagram showing query and response between application layers.

Architecture Implementation

The data access service was developed with the idea that each data source to be supported could be hooked in to the service through modularization. Each data source module is an abstraction of tabularly formatted data that could be queried using a generic syntax. The actual storage technology doesn't need to represent the data this way, so long as the abstraction represents them this way. For example, an abstraction of a relational model could be a set of tabular views into the data accessible through the source's module. For DataZoo the implementation was relatively simple, since the current back-end stores each dataset in a MySQL table. For another project we support, the module abstracted a set of XML files as a table, one record per file. In this way we are now able to query two relatively disparate data sources using a similar syntax and client-server protocol. Additional development was done in order to re-design an existing plotting web service to work with the data access service. The plot service now takes the unique identifier created during a query, along with a set of plotting parameters, and creates a graphic using the access resource of the service to retrieve the data result. It provides the capability within our architecture to create interfaces for querying and plotting data across multiple sources using a single client interface.

Summary

The development of the data access service has helped us move toward providing a common interface for interacting with data that require varying representative models. The asynchronous character of the service provides a stable client-server interaction independent of other application-specific constraints, like web server timeouts. And the stored data result simplifies post-query interactions like viewing, plotting, downloading and re-formatting. We look forward to hearing and learning from other sites' experiences and approaches to scaling issues associated with data access.

Dataset Attributes and the LTER IMC: A First Step

edit

Corinna Gries (NTL), Wade Sheldon (GCE), Karen Baker (CCE, PAL)

A virtual water cooler was held 1-2 Nov 2010 to initiate an LTER IMC process of discussing attribute labels and descriptions. Dataset attribute labels play an essential role in the discovery and integration of data as well as in preparation for mapping to ontologies. The lack of attribute standardization in LTER is being recognized as a major impediment to data discovery as well as to modeling efforts at sites and to cross site synthetic data analyses.

Two existing attribute labeling systems were discussed as examples of different approaches and degrees of standardization. USEPA's STORET code system combines parameter, medium, method, units, etc., within a five digit code number. This rigid approach has led to some problems like force fitting and a proliferation of extensions to the system to fit local needs. CUAHSI on the other hand allows for flexible combinations of some standardized concepts (attribute name, unit, sample medium) with more free form descriptions (methods). The drawback here is that sample methods need to be organized and evaluated by users prior to data synthesis.

Having learned some hard lessons about under-estimating the scope of work involved in standardization (e.g with the Ecological Metadata Language efforts and EcoTrends), it is clear that this undertaking – an information science issue involving classification, category formation, indexing, vetting, managing, and updating over time – represents a large, concerted effort. The goal is broader than a one-time task of standardization; it's a complex community effort involving both practices and tools for developing an ongoing process for attribute standardization.

The first steps include articulating the issues

- Prerequisites to starting the process
- Thoughts on scope and coordination of design, development, deployment, and enactment efforts
- Models of deployment and enactment (e.g. static versus dynamic vocabularies)
- How to include sites, synthesis projects, PASTA, and end-user input
- How to incorporate LTER synthesis efforts as part of the process
- Resource issues (proposals, partners, multi-site LTER supplement efforts)

The next steps include

- Evaluate current practices and consider Best Practices
- Review LTER projects ClimDB and Unit Registry as case exampled
- Consider additional models, e.g. SEADATANET (reference)
- Consider intersections with LTER controlled vocabulary
- Survey LTER site attribute models
- Name and form a working group

An initial step has already been taken by holding an LTER IMC all-site virtual water cooler discussion and presenting the topic for consideration by the LTER Information Management Committee. It is the first step in the process of gathering information and identifying a team of individuals interested in defining the issue. The VTC-initiated discussion revealed a wide range of reactions, with some sites considering this yet another added task and other sites expressing clear visions of how such an effort would be beneficial to site, network, and community level work.

References

SEADATANET. Pan-European infrastructure for Ocean & Marine Data Management and Brisitsh Oceanographic Data Centre (BODC). (http://seadatanet.maris2.nl/v_bodc_vocab/welcome.aspx)

Note on Category Formation

edit

Karen Baker (CCE, PAL)

New categories that arise within a community are noteworthy because they represent ways of organizing information and reflect current community understanding. Categories may initially be understood only vaguely but may become useful, partly because their definitions emerge somewhat naturally over time. The Ocean Informatics team that works with PAL and CCE LTER has found it useful to work with two categories that exist within the LTER community – 'signature' and 'core' – and thereby work on distinguishing the two during this time of category formation in the broader network. Of course, their definitions will likely change or perhaps sharpen over time.

On site websites and within data catalogs, LTER datasets are being tagged as 'signature' and 'core' datasets. In terms of CCE and PAL use, the 'signature' datasets are those that represent a time-series that is long enough to be significant for that type of data; the 'core' datasets are those that focus on LTER themes. These two categories are recognized explicitly to be non-exclusive; that is, a site dataset may be in one, both, or neither category.

Currently, two uses of the 'signature' and 'core' categories are illustrated on the CCE site data page (http://cce.lternet.edu/data/). First, representative examples of signature and core datasets are presented. These serve as both an introduction to site data and as an access point to the data system. Under each category, a graph is presented for the generalist interested in browsing and a dataset link into the data catalog aids the more engaged participant ready to explore. This data presentation design evolved from a collaborative effort by a site scientist, graduate students, and information managers to review the project website.

The second use of these categories highlights the system functionalities built around keywords. As our information system architecture is designed to accommodate multiple sets of keywords, a set of nine keywords has been created for LTER-specific categorization. This set contains the original 5 LTER core science themes (disturbance patterns, movement of inorganic matter, movement of organic matter, population studies,primary production), three additional site-designated core keywords (education/outreach, information management, and social science) as well as the recent addition of 'signature'. For the data presentation discussed above, it has been important to convey that representative examples are given rather than all datasets in each category. In the data catalog, however, all the datasets identified by keywords are available to be retrieved.

The definition of a particular category or set of categories may change over time. With different development cycle times at different levels, having loosely-defined categories or categories 'in formation' at the community-level avoids 'premature standardization' and does not preclude their definition and immediate use at the site-level. Site-level changes can be made quickly, while it takes time for development of community interest and consensus. Though a site must stay in synch with definitions at the community-level, the PAL and CCE use of 'signature' and 'core' categories gives us experience that may inform the development of these or related categories in broader semantic arenas. Our local work may be seen as contributing to network-wide category formation.

Using the OBOE Ontology to Describe Dataset Attributes

edit

Margaret O'Brien (SBC)

Discussions have begun in the Network on the need to standardize dataset attributes where possible. This activity is very important, since the attribute description is what dataset users will focus on, and is probably the most important part of the dataset when it comes to enabling integration. As we develop attributes, we should also consider linking them to our two emerging registries, Units and the Controlled Vocabulary. One way to accomplish such a linkage is through an ontology, a technology which is quickly becoming the basis of the Semantic Web. As we design our attribute registry, it would be advantageous to keep some features of an ontology in mind.

The power of an ontology comes with the relationships that can be defined. There are simple parent-child relationships, as in a taxonomic tree, (“a copepod *‘is-a’* crustacean”), the periodic table, (“a nitrogen isotope *‘is-a’* element”), or ad hoc relationships like “a lake *‘is-a’* water body”. Other more subtle relationships can also be expressed, such as “a tree branch *‘is-part-of’* a tree”. OBOE stands for “Extensible Ontology for Observations”, and by making the observation central, it can be used for annotations at the attribute level. Following is a short introduction to one feature of the OBOE ontology which LTER could consider as we develop dataset attribute descriptions.

Dataset attributes can be defined as OBOE “Measurements” which have these four basic components: Entity, Characteristic, Standard, and Protocol.

1. Entity

The thing that was observed by the measurement. In the case of LTER, the Entity will often be one of the controlled vocabulary keywords (Fig. 1). An ontology also allows synonyms; in this figure, "silicic acid" and "silicate" are equivalent.

- Substances: ammonium, nitrate, bicarbonate, antimony, carbon
- Habitats: aquatic ecosystem, forest, benthic, basins, arctic, clearcut
- Organisms or their parts: bacteria, bark, beetle, arthropod, ascomycete, zooplankton, leaf
- Concepts or processes: biodiversity, atmospheric deposition, carbon cycling, primary production

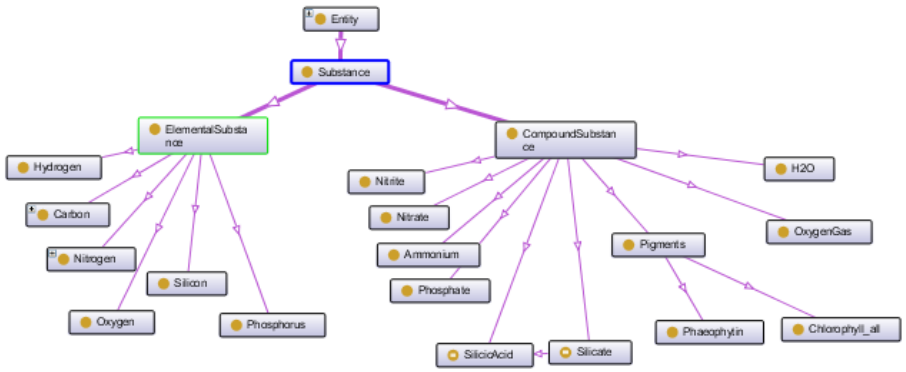


Figure 1. Graph of an OBOE Entity tree for SBC substances.

2. Characteristic

A property of the Entity that can be measured. Characteristics might be dimensionless, like a name or a type (Fig 2). Physical characteristics (like amount, length or concentration) will have dimensions, and the characteristic will be tied to allowable units (see Standard, below).

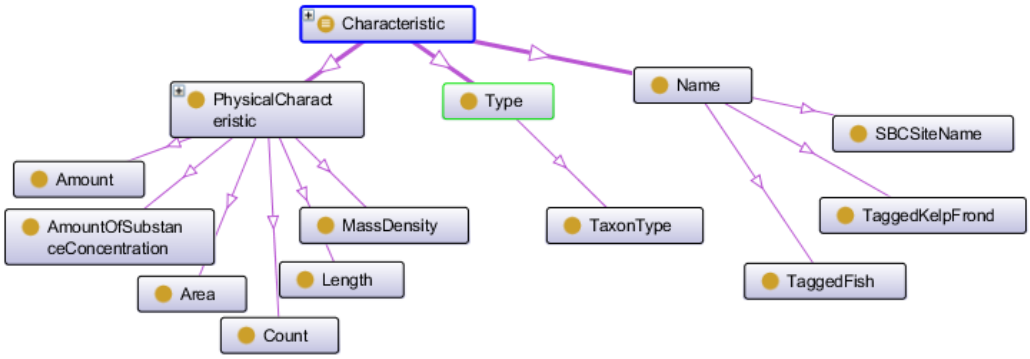
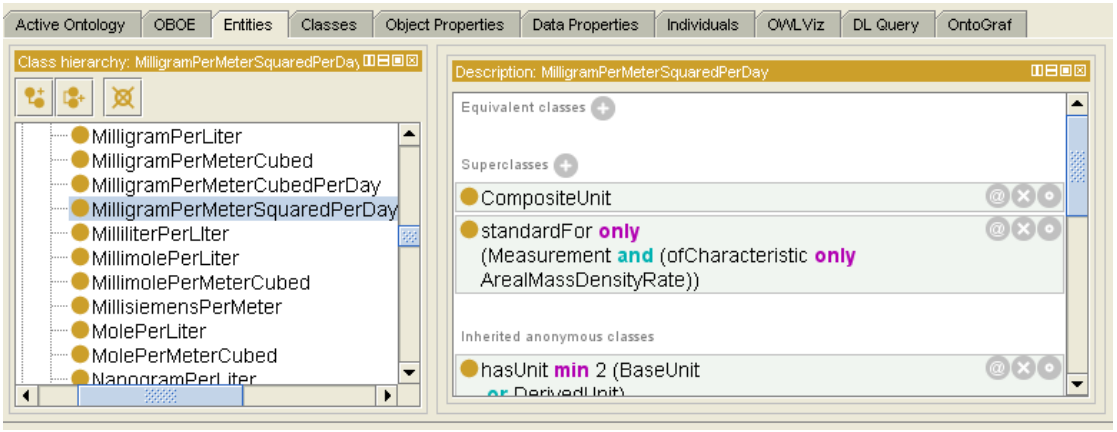


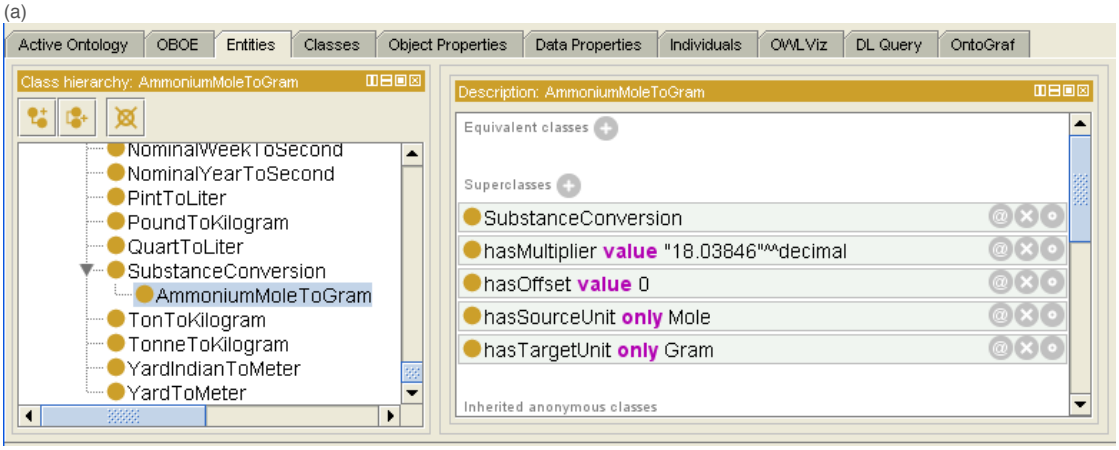
Figure 2. Graph of an OBOE Characteristics tree.

3. Standard

A reference for comparing measurements, e.g., from a dictionary of units, place names or taxa. Some measurements may have more than one standard, for example, a measurement of Ammonium Concentration may allow many choices for unit. The OBOE core ontology imported the units from the LTER Unit Dictionary in 2010 and added their dimensions. Other units may be added as needed. There is also a mechanism to record conversions between units, such as for ammonium from amount (moles) to (mass) grams (Fig 3).



(a)



(b)

Figure 3. Definition of the standard (in this case a unit) MilligramPerMeterSquaredPerDay (a), and a definition of the conversion between a substance, Ammonium, in moles to grams (b).

4. Protocol

The prescribed method for obtaining the measurement. A measurement can have only one Protocol. In ontologies, terms can have multiple parent-terms, so a Protocol may belong to multiple trees. In the example, the Protocol for Dissolved Organic Carbon Concentration can be found under a laboratory ("Carlson Lab Protocols"), under "Protocols by Constituent", and "Protocols for Elemental Analysis" (Fig 4).

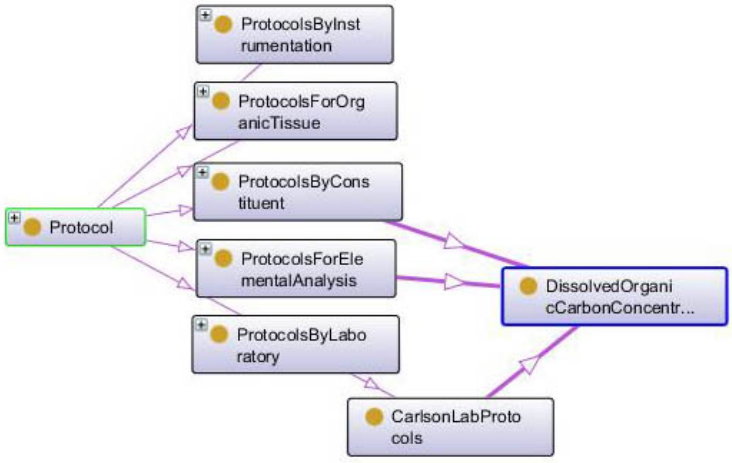


Figure 4. A protocol may belong to multiple trees.

And as one final example, here is the full description of a measurement called "Concentration of Ammonium". In the SBC OBOE extension, it has two subclasses, for fresh water and saline water. If SBC adds another measuremnt of ammonium that required a different protocol (e.g., in anoxic porewater), we could add a third subclass. Since the OBOE measurement description includes many of the same elements as an EML attribute, it is possible to map between the two systems. Not shown here are additional rules for precision and range, which are applied at the most granular level.

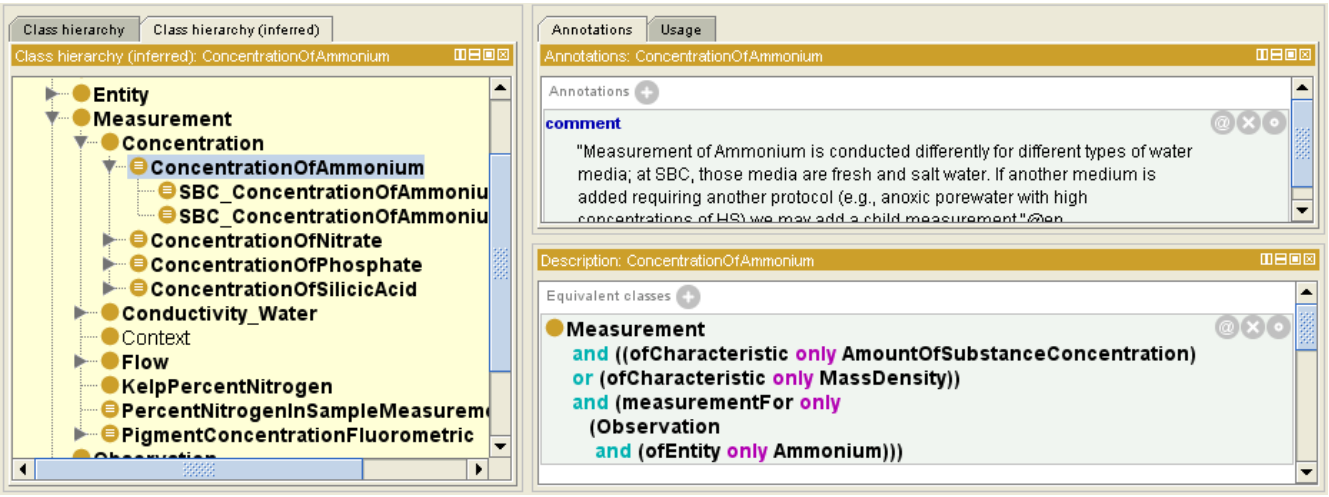


Figure 5. Description of the measurement "Concentration of Ammonium".

Many communities in the life and physical sciences have designed ontologies for their specific needs, and have found semantic technology to be very powerful. As the LTER Network considers the uses and requirements of attribute standardization, it would be advantageous to consider how an ontology might be used to connect our attributes to our other registries. There are several advantages to choosing to work with an extensible ontology such as OBOE. Since 2008, the Semtools project has been working with an LTER site's EML datasets as a use case for an OBOE extension and for tool development. The examples here are from SBC's OBOE extension, and its development has been mindful of the fact that many of the concepts and terms have broad applicability. A related project, the Semantic Observations Network (SONet) has a goal to develop and ratify a community-driven core ontology for representing observational data, and is considering several ontologies and observational models (including OBOE). By working with one or both of these groups, LTER could maximize its success with this emerging technology.

For more information:

- https://semtools.ecoinformatics.org
- https://sonet.ecoinformatics.org
- Ontologies 101. Natalya F. Noy and Deborah L. McGuinness. Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001.
- LTER Attribute Standardization is discussed in another article in this issue (Gries, Fall 2010)

Commentary

Enactment and the Unit Registry

edit

Mason Kortz (CCE,PAL)

The lifecycle of a technical project is often divided into three phases: Design, Development, and Deployment. Science studies researchers have identified a fourth phase that both follows and supports the previous three: Enactment. Enactment is the encouragement of usage of a standard or resource within the target community. Enactment also provides feedback to the developers, which may result in further iterations of the project lifecycle. This article summarizes the Unit Registry enactment activities over the last several months, as well as the benefits provided by this approach.

Summary of Prior Phases

The LTER Unit Working group began a community-based Design phase of the Unit Registry (http://unit.lternet.edu/) in spring 2009. Following the design phase, programmers Mason Kortz and James Conners began the Development phase in summer 2009. The first release of the Unit Registry web service was in fall 2009, followed by a web client with browse, search, and management capabilities in winter 2010. With the completion of the web client, the primary development cycle of the Unit Registry was finished, and in spring 2010 the working group began the Deployment phase. Deployment of the Unit Registry involved migrating web services, clients, content, and documentation to LNO servers, where they would be available to the LTER community. With the software, source code, and documentation deployed, the Unit Registry project proceeded to the Enactment phase.

Cross-Site Visits

The CCE and PAL LTER sites have hosted two cross-site visits in San Diego as part of the Enactment phase of the Unit Registry project: one in May 2010 with Sven Bohm (information manager, KBS LTER), and a second in August 2010 with Ken Ramsey (information manager, JRN LTER) and Justin Jensen (programmer, JRN LTER). The goals of these meetings were numerous: to develop clients that would allow KBS and JRN to use the Registry; to ingest the KBS and JRN custom units into the Registry database; to gain feedback on and make changes to the existing Registry data model, service, and client; and finally, to understand the benefits of a face-to-face enactment meeting so that we could improve future site visits.

The primary focus of the site visits was enactment at a technical level. During both visits, a client was developed that would integrated the Unit Registry web service with the site's existing metadata system. We, as Unit Registry developers, were able to provide guidelines for the development of site clients and rapidly address bugs, missing features, or unclear documentation. The visiting site representatives were able to work with us to design an approach that would leverage the network-level aspects of the Unit Registry while supporting their site needs. By the end of each three-day visit, the site had a functioning client that was integrated with their local site metadata system, and the Unit Registry had been improved for future usage.

In addition to the technical enactment goals, each visit culminated with a discussion about the value of small, face-to-face meetings. While costs prevent meeting individually with every site, the two enactment visits provided two key benefits. First, it allowed a period of rapid design and development for both the sites and the Unit Registry, allowing us to gain experience and make improvements that would facilitate future remote enactment steps. The models that KBS and JRN are using to access the Unit Registry provide concrete working examples for other sites implementing a unit system. Second, the visits demonstrated the viability of site-developed modules providing network-level resources as part of the LTER Network Information System.

Developer/Network Office Visit

In June 2010, the Unit Registry developers visited the LTER Network Office for three days. The scheduled purpose of the visit was to migrate the Registry to an LNO server, completing the Deployment phase of the project. Due to the planning and support provided by the LNO technical team, the software migration took much less time than anticipated. This allowed us to focus on discussions about the integration of the Unit Registry, as well as site-developed web services in general, into the larger LTER Network Information System. This turned out to be an essential step in the enactment not just of the Unit Registry project, but also represented part of the enactment of the Network Information System as a whole.

The Unit Registry developers are responsible for providing a product to the LTER network, but are also clients, or enactors, in the network as well. The three-day visit to LNO gave an opportunity for the NIS developers to engage us in this capacity, providing information and receiving feedback much as we did with KBS and JRN during the site visits in San Diego. The result is that we accomplished both the technical task of migrating the database and the organizational task of learning how the NIS development would impact information managers both as users and developers.

Unit Ingestion Effort

Following the 2010 IMC meeting, Unit Working Group co-chair Linda Powell has been helping sites to submit their custom units to the Unit Registry. Part of the feedback we received during the site visits was that having a common space to address unit standardization would have to occur before the standards-making process could be complete. To meet this goal, we have encouraged all sites to contribute units regardless of compliance with the current Best Practices. The LTER sites have shown great support for this initiative – 17 sites have responded with unit lists, and the remaining sites are working to compile and submit units in the near future.

The Benefits of Targeted Enactment

The Unit Working Group has pursued a policy of targeted enactment for the Unit Registry project. This means that in addition to using broad, open-ended enactment steps – public documentation, demos, and VTCs – we have arranged for and participated in dialogues with individuals or smalls groups to encourage use of the Registry. This type of targeted enactment is expensive,both in terms of funding (for the site and LNO visits) and time (Linda has spent a great deal of time in the last month and a half helping IMs submit their units).

However, there are noticeable benefits to this approach. Targeted enactment allows for rapid response to specific enactors' concerns. Not only does this feedback improve the project as whole, the quick response also improves confidence in the product. Enactors who feel directly engaged are more likely to participate in the project, be it as users, testers, or future developers. Such users are also able to engage other members of the community. Thus the enactment of the project builds on itself and spreads throughout the network. With the Unit Registry, we have already seen sites using and contributing to the Registry who were not part of the site visits, but who are building off the products, both technical and organizational, that resulted from them. As we move forward, we hope that use of the Unit Registry will continue to grow and that the project lifecycle can provide a model for future product development, from Design to Enactment.

Transitions and Comparisons

edit

Lynn Yarmey (PAL,CCE)

I recently returned from a 6-week internship with the National Snow and Ice Data Center (nsidc.org) made available through the Data Curation Masters degree program at the University of Illinois Urbana Champaign and the flexibility of the Ocean Informatics (OI) group at the Scripps Institution of Oceanography. I am now moving on to a recently created position as the Science Data Librarian at Stanford University (yarmey@standford.edu). With school [1], the internship, and my recent and new employment, there is an increased intertwingling [2] amongst the realms of libraries, science and information management for me. All of these changes present an opportunity to reflect on my own local information management experiences with the OI team and the LTER IM community and culture.

Of all the communities I have had the privilege of interacting with over the past busy months, the LTER Information Managers are an amazingly sophisticated and integrative community-of-practice staged for continuing success through cross-fertilization. There is an array of critical community elements in place: informal and formal communication channels (Databits, annual meetings, VTCs, [3]), open attitudes towards cross-site sharing, community working space (working groups, the IM website), awareness and involvement in the internal governance process [4]. These elements allow the LTER IM groups to traverse the distance and language [5] issues that come with distributed collaborative work. Unlike more centralized efforts, the bringing together of such experiences, needs and perspectives of the various sites makes for rich discussions. Also evident, however, is the realization of a more robust set of practices, products and processes that contribute to insights into and development of quality data management and data curation.

Though my location and title are changing, it remains clear to me that my role will in many ways remain that of an information manager. I am looking forward to helping the academic library community address the data landscape illuminated by my work with the Ocean Informatics group, the LTER IMC and others. My title will change to that of librarian, but I am confident that my unique experiences and perspective as a field-based information manager will prove valuable with my new community.

[1] http://databits.lternet.edu/fall-2009/continuing-education-options-information-managers

[2] http://cce.lternet.edu/docs/publication/2006/06spring_databits/intertwingling.pdf

[3] http://databits.lternet.edu/spring-2010/openness-and-transparency-development-lter-network-information-system

[4] http://databits.lternet.edu/fall-2009/imc-governance-working-group

[5] http://databits.lternet.edu/spring-2009/vocabulary-development-tool-community-building

Virtualization, Thin Clients and Cloud Computing: Potential uses in LTER

edit

Virtualization, Thin Clients and Cloud Computing: Potential uses in LTER

John Porter (VCR)

Times are changing – again! The Long-Term Ecological Research (LTER) Network has weathered several major changes in how digital information is managed. The start of LTER in the early 1980s coincided with the start of the movement from mainframe computers to desktop personal computers. Nonetheless, mainframe computers and even punch cards were the staple of early LTER computer use. With the advent of the IBM PC in 1983 and subsequent improvements, the stand-alone PC, with floppy disks, small hard drives and even (for retrieving data from the field) cassette tapes) were increasingly useful for augmenting or replacing mainframe computers, which were themselves being replaced by Unix workstations. With the 1990s came the ability to interconnect the PCs and services such as electronic mail gained currency within the network. Thanks to efforts of then LTERNET Information Manager Rudolf Nottrott, LTERNET was one of the first places where email could move across the proprietary systems used by many federal agencies and the larger research community. The network infrastructure went through an order of magnitude (or more) increase in importance with the advent of the first network browsers in 1993, and the burgeoning World-Wide Web. The mid-1990s saw the introduction of interfaces and tools that let relational databases be directly linked to web sites, leading to a huge increase in the capabilities of web based systems for managing information. By 2000 you started to see Content Management Systems (CMS) used to automate

the nuts-and-bolts of web site development. However, we are about to go through another major change driven by increasing virtualization and advances in cloud computing.

I'll start off talking about some specific changes I've seen at the University of Virginia (UVa) and in collaborative projects. First, UVa is in the process of shutting down all of its public computing laboratories. These labs consisted of PCs and Macs loaded with specific site licensed software required for use in classes. The rationale for having labs was that it would be impractical and time consuming to install, configure and test all the needed software on each individual student's PC or Mac. With the labs gone, how will this continuing need be met? The answer is that they are replacing all the labs with the "UVa Hive" – a system that uses remote-desktop software to link to virtual machines running on a large computing cluster. To use the Hive, one connects to a secure web site in a browser, logs in and connects to a virtual machine using remote desktop viewer. The effect is that one sees a PC booting up (alas, currently running Windows Vista, but likely to be running something better soon), that has a "start menu" pre-loaded with all sorts of useful software (e.g., ArcGIS, statistical packages, Matlab). It is as if you used Remote Desktop on your laptop to connect to your office PC, except with the Hive you are connecting to a PC that exists only as a set of files on a VMWare server. Additionally, all the drives on your local PC are automatically mapped to network drives on the virtual machine, so that data can be easily accessed. The latter is important because when you logout from the virtual machine, it is erased, taking any local files with it.

What are the potential implications of this technology for LTER? Such a technology could be used to make available to LTER researchers a standard suite of analytical tools that are available in a few moments from any network-connected PC. Imagine a system where a single login lets you select a system pre-loaded with LTER-specific tools such as the Morpho metadata editor, the Kepler scientific workflow system, ready-to-go ODBC drivers for connecting applications like Excel to databases, form-based editors for adding data to databases or tools for generating data reports. The system would be available in the matter of about one minute (essentially the boot time of the virtual machine) to any LTER researcher, whether they were at their home computer or not. The computational load placed on the local computer is minimal, because the local computer (be it PC, Mac or Linux-based) is just running a remote desktop viewer. The speed of the resulting analyses depend on the speed of the server running the virtual machine, not the power of the local machine, so that "thin client" computers with very limited speed and memory work just as well as powerful workstations. UVa has set up the virtual machines to be transient (each new virtual machine is generated from a set of master files for a particular type of server, so changes during a session not saved to a network drive are lost), but it is also possible to configure the machines to be persistent, or even to share a single machine among multiple users simultaneously. At the VCR/LTER we currently use the latter model for sharing a virtual machine running software for downloading our wirelessly-connected data loggers. A status display is available to LTER staff both at the field station and at UVa through a remote desktop.

A second tool at UVa that we have been using a frequently is the "Collab" system – a SAKAI-based content management system. SAKAI is a content management system with some similarities to Drupal, Xoops and others. Where it is different is that, whereas most content management systems focus on users being able to add content to a web site, Collab focuses on making it easy for users to generate new web sites. Using web form based-interfaces, a completely new collaborative web site that includes authentication, a place to share files, a shared email address and archive, discussion groups, and calendars can be set up in less than 5 minutes. Once the site is set up, site members can populate it with content, just as if it were a conventional CMS-based web site. Using this tool, individual researchers, including graduate students, can easily set up web sites dedicated to specific projects, however large or small. For example I currently am owner of 20 collab web sites including ones for LTER data submissions, wireless network and data logger configurations, VCR/LTER Graduate Students and many more. Additionally, I am a member of 17 other web sites set up by others, dealing with specific topics or projects.

For LTER the implications of this technology, and related advances in other collaborative tools, are that it simplifies the routine mechanics of sharing data, email, figures and documents among a team of researchers. But what distinguishes it from past technologies is the ease with which such an environment can be created, even by relatively naive users.

Finally, there are all those tools provided by Google and others that are making increasingly sophisticated use of AJAX and other technologies that convert a web browser into a highly capable software platform for word processing, spreadsheets and an increasingly wide array of functions. These browser-based tools can serve as clients for web services or for more specialized and proprietary servers. Google Docs has been heavily used during the drafting stage of several collaborative manuscripts I've been involved in. Its ability to maintain a network accessible "master" document – while also tracking changes and identifying their authors has been a huge boon in producing collaborative manuscripts.

Google Docs is a primary example of "Cloud Computing" where most of the actual work is done by a server over the network, but the look and feel of the application is that it is a local application. The UVa Hive example is a slight variant on cloud computing where the web server is used to download a remote desktop viewer that subsequently runs independently of the web browser. All of the examples above only require "thin clients" – essentially a computer capable of running a modern web browser to operate. Just as the advent of web browsers opened new horizons for what can be used to aid LTER research, cloud computing and virtualization are again expanding what we can do. The trick will be to link the opportunities to researchers in ways that help them be more productive.

News Bits

2010 IMC Annual Meeting, Kellogg Biological Station

edit

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

The 2010 Information Management Committee (IMC) annual meeting was held September 21-24 at the Kellogg Biological Station (KBS). The meeting was well attended with representatives from all 26 sites; about one-third of the sites sent two or more people and the LNO sent several members. In the three-year rotation that also includes meetings in conjunction with the LTER All-Scientist Meeting and the Environmental Information Management Conference (EIMC), this annual meeting affords the IMC an opportunity to meet alone and concentrate on advancing internal working group activities. Plenary sessions were highlighted by visits with Phil Robertson (KBS/LTER Chair), and a videoteleconference with NSF officers Todd Crowl, Nancy Huntly, and Peter McCartney.

The meeting focused on coordination of Site Information Manager (SIM) activities with the development of the Network Information System (NIS). Considerable breakout group time was devoted to three major topics: IMC governance, EML dataset practices and quality, and redesign of network databases and associated web services. Development of the NIS places new demands on IMC members as the expectation for data synthesis and integration is high. Major topics were selected with goals to

1. Clearly describe the role and structure of the IMC and formalize the decision-making process directing our activities
2. Develop quality metrics for EML metadata and data
3. Assure network databases are designed to be useful to and content easily maintained by sites.

Other topics addressed were GIS and the LTERMapS project, the LTER controlled vocabulary, the Units registry, Drupal as an IM development framework, and relational data models for metadata. Strong interest from IMC members led organizers to schedule an optional third half-day to accommodate these working group activities. Following is a summary of the progress of each of these groups.

IMC Governance:

This work group has recommended adoption of a document structure called "Terms of Reference" (ToR) for the IMC and its working groups to describe their membership and activities. ToRs are similar to by-laws, but use a bottom-up approach and are somewhat more informal. At the meeting, the Governance working group assembled feedback on composition and format for the ToR from all IMC members, and began a draft ToR document for the IMC which will be available for comment in early December. Once adopted, active IMC working groups will create their own ToRs in the following months. The IMC also approved IMExec as the

body which will review and endorse proposals on their behalf before submission to LNO for the coming year for NIS Production Workshops, Training, IM release time, and IM travel. IMExec has begun this process with the first round of proposals (which were due November 1). This review process will be evaluated by the IMC at their meeting in September 2011. The Governance working group is chaired by Karen Baker (CCE/PAL), Eda Melendez-Colom (LUQ) and Nicole Kaplan (SGS). (See article this issue.)

EML Best Practices and Congruency Checker:

The original EML Best Practices is undergoing revision to reflect changes in EML and to assure consistency of approach in writing EML across the network. The EML Congruency Checker will be developed in 2011 to provide feedback on completeness and usability of EML datasets submitted to the NIS. Meeting breakout time was used to evaluate the draft EML Best Practices document and consider requirements for the EML Congruency Checker. The draft has been available for comment since September 1 and the comment period ends January 15. Work on the congruency checker is expected to be completed during the first half of 2011, as part of the NIS Data Management Suite. The EML Best Practices and Congruency Checker groups are chaired by Margaret O'Brien (SBC).

Network database redesign and web services:

Network databases such as personnel, the all-site bibliography and site descriptions (SiteDB) are typically maintained secondarily by LTER sites and are not well integrated with site information systems. As the NIS is developed, sites will want to make use of network database content for their local uses, e.g., directly use the network personnel list for site purposes or synchronize site publications with the all-site bibliography. This group has two major tasks: to define use cases for how a site might use Network databases, and to recommend web services mechanisms for accessing information or synchronizing with Network databases. Meeting breakout sessions were used to consider how these network databases could be redesigned to best accommodate both site and network uses. This group has decided to focus early efforts on the personnelDB and identified the need for a product-oriented workshop to further consider web services development. The combined working group is chaired by Mason Kortz CCE/PAL).

GIS/Maps:

The GIS working group met during a breakout session and discussed several topics including its own governance, site training needs, and feedback to the Strategic and Implementation Plan (SIP) on GIS strategies for LTER sites. A GIS survey was prepared and has subsequently been sent to all sites. Several ideas for training were identified and a training proposal is under consideration. Adam Skibbe (KNZ) will be the new chair of this working group replacing Theresa Valentine (AND). A sub-group of the GIS working group, LTERMapS, intends to provide a consistent interactive mapping interface for LTER site information through an Internet mapping application which allows users to visualize, search, download, and explore site information. The project used the last day of the meeting to plan a November EML-GIS workshop at the Andrews to develop a product-oriented proposal for LTERMapS phase 2 implementation. Three subgroups were formed: Templates, Front-end and Back-end in planning for the implementation. Theresa Valentine is the lead contact for the LTERMapS group.

Controlled vocabulary:

The working group considered next steps to build on the list of 650 terms based on widely-used LTER EML keywords with the goal to include at least one LTER keyword in each EML file. A key next step is to consider the "vocabulary management plan" including its governance and procedures for modification. Polytaxonomies, e.g., a series of taxonomies (habitat, LTER core areas) will be developed. Future tools for ingestion of keywords into EML and for searching databases will be considered. A product-oriented workshop proposal is planned for 2011.

Units Registry:

This group reported on the success implementing the Unit Registry at sites during the past year (with ASM funding), including deployment at LNO, and implementation at four sites. As of this writing, the Registry has been populated with units from 18 sites. Their current activities include outreach and training on the registry via VTC and site-site visits, and further adaptation of the Best Practices for Units document to include discussion of non-compliant units, abbreviations and definitions, assigning quantities, and the unit vetting process. The group will also consider how the Unit Registry may inform development of standardized attributes. This group is chaired by Mason Kortz (CCE/PAL) and Linda Powell (FCE). (See article this issue.)

Drupal Environmental Information Management System (DEIMS):

Six LTER sites, one OBFS site, and NBII are collaborating to develop a set of applications in Drupal which will provide the functionality for a basic research site/field station information website (including, personnel, publications, research sites, datasets and metadata). Currently the main foci are an EML module consisting of a simple metadata editor which maintains metadata in Drupal and produces EML documents, and a query application which allows users to retrieve data subsets. The DEIMS group met at KBS to finalize the agenda of their training workshop, to demonstrate some new interfaces, and to share progress populating backend systems at individual sites. Their upcoming tasks include implementing web services for the Units Registry, with a similar plan in mind for LTER Keywords as these develop. The LTER sites working on this project include LUQ, SEV, ARC, PIE, NTL and VCR.

Relational data model comparison.

Relational data models for maintaining site metadata are often a key component of site information systems and exist in varying schema structures throughout the network. Some site IM systems have proven agile enough that their data models evolve with changing needs, while other sites are exploring possible improvements or are planning to adopt a model from another site. As part of their adaptation process, a comparison of existing data models was constructed by M. Gastil-Buhl (MCR) and presented as survey results in a poster and discussion. At the time of the meeting, three examples of mature production models which are part of a larger IM System were presented: the GCE Metabase, the AND Metadata Database, and the CCE/PAL DataZoo. These models continue to undergo improvements, and their content is used to provide services such as web page display and EML generation. A fourth model using the Drupal Content management system is in development (DIEMS) by a group of six sites. The comparison has continued since the meeting with input from additional sites. The benefits of this comparison are 2-fold: it allows sites to understand the most successful features of existing systems as they consider adapting one of the models, and secondly, allows sites to compare their systems' features to other site systems within the network.

Clearly, the IMC has made significant progress on its 2010 projects, and is well placed for work in 2011. We anticipate that in the coming year, we will see a streamlined and accessible Network personnel database and significant development on a keyword dictionary, which will allow us to more efficiently manage diverse tasks and data contributions. We anticipate that reports describing quality and usabililty of datasets will highlight the areas where our IM systems are in need of upgrade, and help to provide justification for additional funding supplements. IM-Exec appreciates the strong interest and dedication of the entire IMC at this year's meeting.

GIS Working Group holds successful workshop at HJ Andrews Experimental Forest

edit

Theresa Valentine (AND)

The GIS Working Group held a successful workshop at the HJ Andrews Experimental Forest. The focus of the workshop was documenting GIS/remote sensing data and converting the documentation into EML (Ecological Metadata Language). 11 people attended on-site, and there were 5 sites represented through video conferencing. The outcome of the workshop will be a document that outlines the different paths to convert GIS metadata to EML (depending on the software versions that are used), and recommendations for changes in the conversion program. The group also worked on the next phase of the LTERMapS project.

Workshop participants, left to right: Eda Melendez-Colom, Yang Xia, Adam Skibbe, Hope Humphries, Theresa Valentine, John Van Castle, Jamie Hollingsworth, Barbara Nolen, Suzanne Sippel, Jonathan Walsh and Aaron Stephenson in person and on video clockwize from upper left: John Porter, Robert Flynn, Mike Friggens, Wade Sheldon, Travis Douce, and Margaret O'Brien.

On Site with TFRI

edit

-Benjamin Leinfelder (NCEAS)

Though long-time collaborators, my recent two-week trip to Taiwan reified the collaborative momentum that has grown between NCEAS[1] and TFRI during the past year. With an increasing focus on international information management, the ecoinformatics team at the Taiwan Forestry Research Institute - headed by Dr Chau-Chin Lin - remains a prominent advocate for robust data stewardship practices. I facilitated a series of hands-on tutorials and discussions that included topics on advanced Metacat deployment techniques, new Morpho features, extended applications of EML, and a primary focus on integrating Kepler as a crucial component in ecological data collection, management and analysis.

TFRI supports a network of research sites across the country with sensor arrays for collecting meteorological data, still images, and audio and video recording. Equipment deployed near the ecological pond at the Lienhuachih research station (figure 1) continuously streams audio to a central server where it will be used as raw biodiversity data. Similarly, a camera is trained on a patch of foliage to monitor phenological events in the preserve. The site serves as a proofing ground where Kepler will be employed for realtime quality assurance, data analysis, and archiving. Site-specific initiatives such as this dovetail nicely with existing techniques that NCEAS has pioneered on the REAP project[2] where Kepler workflows are used to monitor and manage sensor data collection from Data Turbine servers[3].

The usual sensor-in-the-field hurdles are not absent here: high humidity, varied terrain, frequent severe weather, unreliable wireless networks, finite battery power, etc. But a strong commitment to working through these physical barriers is palpable and evidenced by the zeal with which the TFRI team approaches their projects. They are well positioned to transition from being superusers of the KNB ecoinformatics software stack to being developers themselves, in true collaborative fashion. Moreover, they are able to provide concrete use-cases that challenge the current state of our technology and motivate enhancements to usability, interoperability, and quality - a boon for existing and emerging [international] cooperative data networks.

Figure 1. Dr Yu-Huang Wang describing the audio recording equipment in the field at Lienhuachih

[1] National Center for Ecological Analysis and Synthesis (http://nceas.ucsb.edu/)
[2] Realtime Environment for Analytical Processing (http://reap.ecoinformatics.org/)
[3] A server for realtime streaming data (http://www.dataturbine.org/)

Good Tools And Programs

Making the Work Flow with Kepler

edit

John Porter (VCR), Chau-Chin Lin (TERN), Jennifer Holm (LUQ), Ben Leinfelder (NCEAS)

During the summer of 2010 as part of the “Second Analytical Workshop on Dynamic Plot Application and Tool Design” in Kuala Lumpur, Malaysia, we had an opportunity to employ the Kepler scientific workflow tool in an authentic research context. The workshop brought together experts on tropical forests and ecological informatics to use innovative computational tools to examine how the biodiversity and spatial structure of forests change between locations around the world (see the Fall 2010 LTER Network News for more information about the workshop). Here we will focus on how we used Kepler and the challenges and opportunities it provided.

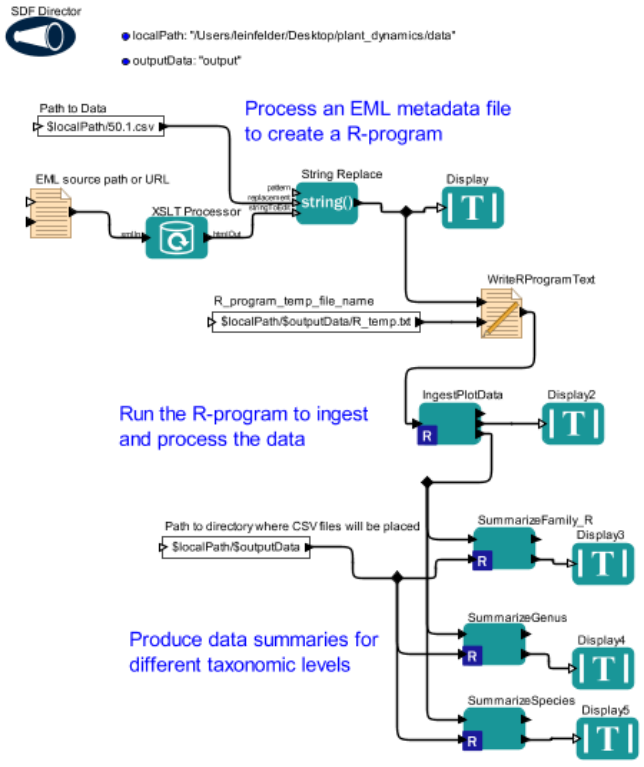
The challenge we faced in the workshop was the integration of somewhat heterogeneous datasets from mapped forest plots. The forest plots used in this study are long-term permanent plots set up in locations from North America to Central America and many parts of Asia. Each site is managed by different scientists leading to the discrepancies in data gathering between locales. The data had some similarities (all had taxonomic designations, tree measurements and coordinates), but differed in the detail of taxonomic data, how the status of stems was designated (live, dead, main, secondary), the way it was structured (some were in a single table, others in multiple tables) and the names of fields. These data needed to be ingested, converted into standard forms, processed statistically to produce new summary data structures and analyzed. The following lessons may apply for any group that will work with dissimilar and large datasets from around the globe.

We chose to accomplish these tasks using the Kepler workflow system (http://seek.ecoinformatics.org/). Kepler was selected because it was freely available, had good support for Ecological Metadata Language (EML) and for the “R” statistical language. Kepler workflows appear as a set of interconnected boxes (“actors” in Kepler terminology), with each actor having one or more “ports” where connections can be made. Kepler has actors that provide automated EML data ingestion, execution of “R” programs, XML-stylesheet processing, text manipulation capabilities, and displays, among many other functions. For the RExpression actor, ports equate to scalar and vector elements that share the same name as the port. Ports can be used to output many forms of data, but not all of them are specified in the pull-down list, so it may require some experimentation to find the right output type to match the input of other actors if Kepler cannot automatically determine the appropriate data type. In Kepler a “director” controls the order of operations for actors. For analytical functions, as compared to modeling, this is almost always the ”SDF” (synchronous data flow) director, with the number of iterations set to 1.

During the course of the workflow development we needed to address several issues. The first issue involved using the EML actor to ingest very large data files caused Kepler to crash with a Java stack overflow exception. Because actors intercommunicate using data “tokens” in memory, large data files in excess of 35 MB would cause the workflow to fail after protracted processing. To address this problem, we used the XML stylesheet processor actor to transform the EML document directly into an R script. The large data files were parsed and loaded dynamically from a specified location on disk and saved as an R workspace. Subsequent R actors were able to load this workspace without the need to pass the voluminous data values via Kepler ports. An additional issue was that error reporting in the RExpression actor was rudimentary; workflow execution might fail, but Kepler would not provide specific error messages from R regarding the nature of the Therefore, we typically wrote and debugged R code outside Kepler, before adding it to the workflow. The active Kepler developer community has been informed of these issues and always encourages user feedback, so hopefully future versions of Kepler will resolve these problems.

After addressing these issues Kepler workflows worked quite well for ingesting and processing the data (Figure 1). The workflows successfully processed the data while also effectively communicating the analysis to the workshop group in a portable format. Workflows originally created in Virginia were revised in Taiwan and run in Malaysia. As relatively new Kepler users, we did find that there was a significant learning curve. Early workflows required about 4-times to create as might be spent “manually” performing processing. For later workflows the time required dropped dramatically, primarily because Kepler made it easy to adapt and reuse workflows. As a graphical tool, Kepler is advantageous in that it organizes the entire analysis and allows researchers to see both the high-level process as well as each detailed step. Workflow parameters can easily customize an analysis while text annotations can highlight pertinent notes about the workflow before it is distributed as a single, self-contained file. These features are especially important for complex analyses being developed by a large and spatially distributed working groups where clear communication and rapid development is essential to the collaboration.

Figure 1: Kepler Workflow for ingesting and processing data from a forest plot.



SchemaSpy: No dust will collect on this Database Documentation.

edit

M.Gastil-Buhl (MCR)

SchemaSpy is a very useful program for creating database documentation from the database itself, and because it does so automatically, it is easy to keep such documents current. Of the six products I trialed, not only is this program the most universal but also creates the most useful description of database structure. DBScribe, described in Databits (W.Sheldon, Spring 2009), although a good product, only runs on PCs and cannot connect to a remote PostgreSQL host.

Recently I ported some of the databases Wade Sheldon wrote for GCE-LTER from SQL-Server to PostgreSQL. Those familiar with Microsoft SQL-Server Management Studio know how convenient it is to generate Entity-Relationship Diagrams (ERD) and MySQL Workbench offers a similar feature. SchemaSpy works for all the rDBMS used by LTER sites (MySQL, Oracle, MS SQL-Server, and PostgreSQL). It is independent of the OS (PC, Mac, Linux) because it is written in Java. It can access a database on the local host or a remote host. It runs on the command line so it can be invoked automatically at regular intervals. (There is also a GUI built for it which I did not test.) To summarize: SchemaSpy creates (or refreshes) the documentation with a single command; that part is not interactive. The documentation itself is an interactive webpage. The features I describe below are within that webpage.

Figure 1. The documentation is published as a tree of html pages which appear as tabs.

The tables in the ERD are mapped links to expand or contract scope to explore the data model of a whole database, just one schema of the database, or selected tables. The view of part of the ERD can be interactively expanded to show just the neighboring parent and child tables or all relations within two generations either side at once with toggle. SchemaSpy automatically places parent tables to the left and child tables to the right of the selected table. The foreign key relationships between tables point directly to the referenced columns (shown as rows in the box for each table). In most ERD-drawing tools the point where this line connects is not automatically aligned with specific columns. Figure 2, below, is a screenhot; the SchemaSpy homepage offers a live example. Sharp eyes will note the row counts for each table. As this database was just ported, no data has been entered yet. Because SchemaSpy is easy to refresh, I did not wait until the final version of the port is completed. Even small changes, such as converting the SQL-Server bit data type to boolean in PostgreSQL are worth updating in the documentation, especially when development is a collaboration between sites. Figure 2 shows just a subset of the tables in the GCE implementation of ProjectDB; compare to the complete ERD shown in Sheldon & Carpenter, this issue. A disadvantage of diagrams drawn by SchemaSpy is that they are optimized for interactive browsing rather than creating a graphic that fits a whole ERD into one page. (For that I use another graphviz-based tool.)

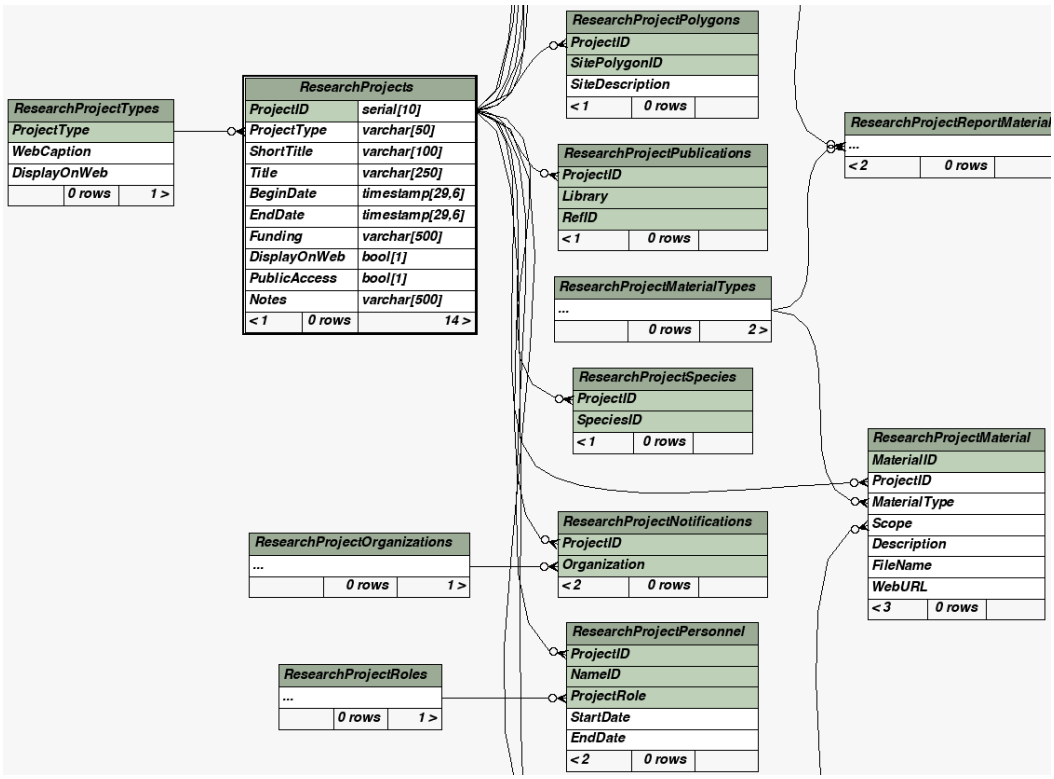


Figure 2. Entity-Relationship Diagram of a subset of the GCE implementation of ProjectDB as drawn by SchemaSpy. Primary key columns are green.

Complete and accessible documentation is necessary, but not sufficient; it must be maintainable. How large a fraction of our work time can we afford to spend documenting and updating the documentation? Time invested depreciates as the continually developing systems diverge from static documentation. Recently, during the VTC with our NSF program manager at the September IMC meeting, we were reminded of the value of "sustainable software". A necessary part of that is maintaining useful documentation that is both technically rich as well as meaningful to non-technical stakeholders. The easier it is to maintain documentation, the more likely it will be kept up to date by a busy Information Manager. SchemaSpy can be re-run with one command to update the documentation of a database, using only the database itself.

Further Information:

- SchemaSpy homepage on sourceforge: http://schemaspy.sourceforge.net/
- Described in 'Java Power Tools' by O'Reilly, 2008, Chapter 30 'Automatically Generating Technical Documentation'.
- Dependency: Graphviz (optional but the most useful part.)
- Summary on Freshmeat:
  - License: LGPL "Lesser" General Public License [2]
  - OS Independent because implemented in Java

- "SchemaSpy analyzes database metadata to reverse engineer dynamic Entity Relationship (ER) diagrams. It works with just about any JDBC-compliant database and can identify Ruby on Rails style databases, as well as other implied relationships."

- GUI version also in Java.

Good Reads

Evolution of Collaboration in Ecology

edit

Karen Baker (PAL,CCE) and Eda Melendez-Colom (LUQ)

Review: *'Evolution of Collaboration in Ecology' by W.Michener and R.Waide. In Scientific Collaboration on the Internet. G.M.Olson, A.Zimmerman, and N.Bos (Eds). The MIT Press, Cambridge, 2008.* ISBN-10:0-262-15120-0

This book chapter by Michener and Waide describes the Long-Term Ecological Research program and "how collaboration has evolved in a like-minded community of ecological scientists". The 20 chapters in the book are grouped into those relating to the notion of collaboratories (collaborative science at-a-distance) and those about the sciences that are subgrouped into sections on physical sciences, biological and health sciences, earth and environmental sciences as well as the developing world. The earth and environmental section includes chapters on:

- NCEAS (Ecology Transformed)
- LTER (Collaboration in Ecology)
- GEON (Organizing for Multidisciplinary Collaboration)
- NEESgrid (Cyberinfrastruture Development)

The LTER story is told by interleaving LTER history with general impacts of agency funding for LTER as well as NCEAS, SEEK, KNB, and NEON as well as with increased communication and coordination. Six metrics of success of LTER and nine lessons learned about collaboration are presented in this chapter. The nine lessons learned in the LTER evolving culture of collaboration are summarized as:

1. Establish or identify a common vision and common objectives
2. Provide support for face-to-face communication
3. Invest in developing and adopting standards
4. Support cyberinfrastructure and information management
5. Be flexible and engage stakeholders in the process
6. Recognize the value of incentives and oversight
7. Look beyond your normal comfort zone for ideas and collaborators
8. Learn from your predecessors
9. Leverage

The examples of LTER collaboration presented will be of interest to information managers who have gained experience with the work of articulation and openness, providing communication across multiple levels, infrastructuring at the site-network interface, engaging participants within the network, and taking long-term - even historical - views of a multiplicity of successes and failures.

A set of metrics are also given as evidence of the LTER as "a major scientific success story". Listed as one of the lessons learned in association with cyberinfrastruture, the development of the role of information management at both the site and network levels is not included as one of the metrics for LTER success. There are many kinds of metrics including quantitative measures as well as identifiable performance milestones. The notion of collaboration and the concept of metrics of success provide, therefore, an important prompt for all participants to mull over their own LTER experiences while asking: what is uniquely effective about the LTER Network configuration for synthetic scientific research, and how can it be captured as a milestone?

Healthy Tensions, Challenges in Achieving Data Sharing

edit

M. Gastil-Buhl (MCR)

Review: *'Infrastructuring Ecology: Challenges in Achieving Data Sharing' by Karen S. Baker and Florence Millerand, Chapter 6 in Parker, John N., Penders, Bart, and Niki Vermeulen (Eds.) October 2010. 'Collaboration in the New Life Sciences.' London: Ashgate. ISBN: 978-0-7546-7870-0*

"Ecological data remain closely tied to traditional disciplinary knowledge .... Yet data initiatives today frequently focus on data reuse ... outside the domain ...." The authors explore "as to whether models of data sharing can be borrowed and imported with equal success in ... the environmental sciences characterized by research data and data practices that are highly heterogeneous and complex." Using the LTER as a case study, the authors demonstrate how ecological data can be more complex and diverse than physical data or data within one constrained domain of biology. The authors observe that "interpreting datasets of diverse types brought together across multiple scales can be a research project in and of itself, a tacit underappreciated part of the scientific process of knowledge building. That is, the mechanics of assembling data in a central location differs from the frequently iterative work of processing and reformatting data in order to be able to interpret *and to evaluate* an integrated result." They go on to deconstruct the data lifescycle in two contexts, internal use and reuse external to its original context. The authors analyze concrete examples of past efforts in data integration and data sharing, and posit lessons to inform our current practices.

Selected highlights:

"Contemporary cyberinfrastructure initiatives are throwing light on data and data practices in the sciences in two principal ways: first, in promoting larger-scale scientific collaboration and second, in making new arrangements for data sharing and more formal digital data publication."

"In formalizing the data analysis and the data curation subcycles (of the data lifecycle), metadata ... make visible and organize knowledge currently held tacitly." These subcycles are "iterative processes comprised of planned tasks and ... unanticipated irregularities". There are "healthy tensions involving local context-sensitive impulses to accommodate and remote curation-driven impulses to standardize data differences together with the mix of analysis-intensive research impulses to learn from anomalies and data-intensive synthetic efforts to learn from patterns."

Summary:

This book chapter is worth an Information Manager's time to read. Although this study used ethnographic methods, for us as Information Managers it is an insightful introspection. The longitudinal perspective in time helped me, as a relative newcomer, to better understand our current challenges. The next proposal or IM plan I write I will be able to cite an up-to-date reference for the statement 'Information Managers have studied the challenges to achieving data sharing and can build upon those lessons.'

FAQ

Is there a preferred stable URL for the EML schema?

edit

Mark Servilla (LNO), Margaret O'Brien (SBC) and Wade Sheldon (GCE)

Question:

Is there a preferred stable URL for the EML schema to put at the top of my EML documents?

Answer:

The network office (LNO) now provides these URLs:

- **EML**
  - http://nis.lternet.edu/schemas/EML/eml-2.0.0/eml.xsd
  - http://nis.lternet.edu/schemas/EML/eml-2.0.1/eml.xsd
  - http://nis.lternet.edu/schemas/EML/eml-2.1.0/eml.xsd
- **LTER ProjectDB**
  - http://nis.lternet.edu/schemas/EML/lter-project-eml-2.1.0/lter-project.xsd

Put the URL for your document's EML version in the <eml> element the top of each EML file. This allows it to be validated by applications. When the schemaLocation URL is defined, oXygen will become aware of the schema and will mark any schema-invalid xml so it is easier to catch errors. There are some caveats when using the EML 2.0.1 schema; for further explanation see the Metadata section of the IMC website at: http://intranet.lternet.edu/im/im_practices/metadata/eml_versions_validation

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <eml:eml xmlns:eml="eml://ecoinformatics.org/eml-2.1.0"
3      xmlns:stmml="http://www.xml-cml.org/schema/stmml-1.1"
4      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
5      packageId="knb-lter-mcr.31.24" scope="system" system="knb"
6      xsi:schemaLocation="eml://ecoinformatics.org/eml-2.1.0 http://nis.lternet.edu/schemas/EML/eml-2.1.0/eml.xsd">
7      <access authSystem="knb" order="allowFirst" scope="document">
```

Figure 1. The top of an EML document being edited in oXygen, showing the xsd (line 7)

Calendar

Events 2010-2011 Winter & Spring

edit

Event: Digital Curation Conference (DCC)

Location: Chicago, Illinois, USA
Dates: December 6-8, 2010
Web: http://www.dcc.ac.uk/events/conferences/6th-international-digital-curati...

Scientists, researchers and scholars generate increasingly vast amounts of digital data, with further investment in digitization and purchase of digital content and information. The scientific record and the documentary heritage created in digital form are at risk from technology obsolescence, from the fragility of digital media, and from lack of the basics of good practice, such as adequate documentation for the data.

Event: AGU

Location: San Franscisco, California, USA
Dates: Dec 13-17, 2010
Web: http://www.agu.org/meetings/fm10/

This year AGU has sessions on informatics, and Infrastructure for Sciences, and Sensor Networks. Several LTER IMs are presenting, including Corinna.

Event: Hawaii International Conference on System Sciences (HICSS)

Location: Kauai, Hawaii, USA
Dates: January 4-7, 2011
Web: http://www.hicss.hawaii.edu/

Since 1968 the Hawaii International Conference on System Sciences (HICSS) has become a forum for the substantive interchange of ideas in all areas of information systems and technology. The objective of the conference is to provide a unique environment in which researchers and practitioners in the information, computer and system sciences can frankly exchange and discuss their research ideas, techniques and applications. Registration is limited.

Event: Computer Supported Cooperative Work Conference

Location: Savannah, Georgia, USA
Dates: 6-10 February, 2011
Web: http://www.cscw2010.org

The conference brings together top researchers and practitioners who are interested in both the technical and social aspects of collaboration. In recent years the conference has moved beyond traditional "work" to include the broader issues of how we play, socialize, and compete - all forms of collaborative activity that are now mediated by technologies. As more and more people in all regions of the globe are able to interact online we are rapidly moving toward a Computer Supported Cooperative World. Appropriate topic areas for CSCW include all contexts in which technology is used to mediate human activities such as communication, coordination, cooperation, competition, entertainment, education, medicine, art, and music.

Event: EIMC - Environmental Information Management Conference

Location: Santa Barbara, California, USA
Dates: TBC, but 2nd half September, 2011
Web: none yet

The annual IMC meeting will be scheduled the day before the EIMC

Event: American Society for Information Science and Technology

Location: New Orleans, Louisiana, USA
Dates: October 7-12, 2011
Web: http://www.asis.org/conferences.html

Since 1937, the American Society for Information Science and Technology (ASIS&T) has been the society for information professionals leading the search for new and better theories, techniques, and technologies to improve access to information. ASIS&T brings together diverse streams of knowledge, focusing disparate approaches into novel solutions to common problems. ASIS&T bridges the gaps not only between disciplines, but also between the research that drives and the practices that sustain new developments. ASIS&T counts among its membership some 4,000 information specialists from such fields as computer science, linguistics, management, librarianship, engineering, law, medicine, chemistry, and education.