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

Fall 2008

With at least one article in six of the seven categories, there's plenty to read in this issue of the LTER's Databits. This Featured Articles section contains an introductory technology guide, a review of this last year's introduction of video-conferencing practices into the LTER IM community and more. Read two reviews of papers centered on data management practices in the Good Reads section. The News section contains an update and review of a current LTER IM project, and in Commentary read an overview of a course taught by the well-known Edward Tufte. Also, there are some new useful applications reviewed in the Good Tools and Programs sections. And don't forget to take a look at the Calendar in this issue for a list of upcoming event dates. Enjoy!

Featured Articles

Getting Started With Web Services
Telling the Story behind the Photos

Commentary

14 Quotes on "Simplicity" by the Greatest Thinkers of All Time.
Whirlwind Tour of Digital Curation in the UK
IMs from the file and server rooms, and to the home office and beyond: Reporting our experiences with video-teleconferencing...

News Bits

PROJECTDB – Planning and Development of a Collaborative Programming Effort

Good Tools And Programs

Creating Archival Documents
MySQL Workbench: A Visual database design tool
Mapping your mind with FreeMind

Good Reads

Enabling long-term oceanographic research: Changing data practices, information management strategies and informatics
Disputed Definitions
Digital Data Practices and the Long Term Ecological Research Program Growing Global

Calendar

Events

Featured Articles

Getting Started With Web Services

edit

Mason Kortz (PAL/CCE)

This article provides basic definitions and resources for developers starting to work with web services. It is an attempt both to organize the concepts I've learned over the last six months and to make the introduction to web services simpler for other LTER developers. It is far from a comprehensive guide to the development and use of web services. I encourage anyone with additional links, resources, or thoughts on web services to add to this article and add to our collective body of knowledge. To this end, I have started a thread on the LTER IM forum (link TBD) where additions to this article can be collected.

Terms and Definitions

These terms cover the basic language of web services. The definitions below focus on these terms as they are used in the context of web services; many of them have meaning in broader contexts. A far more complete (and more technical) set of definitions can be found at the W3C Web Services Glossary.

Service: An abstract interface to the functionality of an application through a set of standardized protocols. Services often provide machine-to-machine interface capabilities using a client-server model, but services may also communicate locally (i.e. communication between operating system components).

Web Service: A service that allows requests and responses to be exchanged over the Internet. Although web services do not, by definition, have to use the HTTP protocol, this is usually implied. In practical use, the term 'web service' is generally limited to services with responses intended to be used programmatically, rather than formatted for human readability by a browser.

Consumer: An application that sends requests to a web service and process the responses.

Provider: An application that handles web service requests and provides responses.

SOA: Service Oriented Architecture. A modular system architecture in which applications provide functionality via autonomous, discoverable services. This allows complex processes to be performed by combining these services, rather that creating an application for each process.

SOAP: Simple Object Access Protocol. An XML protocol for exchanging messages between computers. These messages can be used to access web services in function-call-like manner. The consumer sends an XML request, which contains a function name and possibly parameters. The provider processes the request (i.e. calls the function) and returns an XML response. Both the request and the response are structured and strongly typed, meaning that each parameter and return value is given a type, such as integer, string, or array. New types may be defined by the combination of existing types, allowing complex data structures to be passed as parameters and return values.

REST: Representational State Transfer. An architecture for distributed hypermedia systems. In terms of web services, this refers to a service that accepts an HTTP request in the form of a subject (URI) and an action (GET, POST, PUT, or DELETE) and returns a representation of that subject in the resulting state. An important feature of a REST service is that the request contains all of the information needed to identify the resource and the state – there is no state information held in cookies or session variables. REST does not specify a structure for the request or response beyond the HTTP specification. Retrieving a static web page is a very basic REST operation.

WSDL: Web Service Description Language. An XML schema allowing the description of web services, including the URI of the web service, the functions it supports, the inputs and outputs of those functions, and the transport language used to communicate with the provider. With these specifications, a consumer application can automatically generate native calls to access the web service – for example, a PHP application could read a WSDL file and construct a PHP object with matching methods that would convert PHP calls into SOAP envelopes, and SOAP responses into PHP data structures. Theoretically, WSDL can be used to describe any web service regardless of transfer format and protocol, but in practice is most often used for SOAP services.

Developer Tools

Many languages have programming interfaces available to interact with web services by constructing SOAP envelopes and automatically prototyping functions based off of WSDL files. Desktop tools are also available for interacting with SOAP services and are very useful for viewing the XML level exchange between the consumer and the provider. REST interfaces and tools are fewer because of the lack of a description schema like WSDL, but various standards and code to support them are emerging.

Programming Interfaces

PHP: PHP4 supports SOAP through the PEAR::SOAP package. PHP5 provides similar support natively if the SOAP option enabled at compile. Both extensions can be used to write SOAP clients and servers. PHP can interface with REST services natively by opening URLs if the *allow_url_fopen* configuration is enabled.

PERL: SOAP and WSDL support are available through the SOAP::Lite package. This allows PERL scripts to act as both clients and servers in SOAP exchanges. REST services can be accessed using the HTTP functionality in the LWP package.

MATLAB: Current versions of MATLAB support SOAP transactions natively.

Python: Support for SOAP transactions and WSDL generation/interpretation are available through the SOAP.py package.

Other Tools

Google REST Describe/Compile – An experimental implementation of the WADL (Web Application Description Language) for defining REST web services. It's still in a demo state, but an interesting look at where REST services may be going.

Mac SOAP client: A lightweight desktop gui for exploring SOAP services.

Generic Web SOAP client: A web interface that provides basic SOAP service access.

Further Reading

There are a great number of excellent articles on web services available online, but separating the wheat from the chaff can be a little frustrating. The following articles were ones I found particularly informative, interesting, and enjoyable to read. For more information, the W3C web services home page and the Wikipedia article on web services are both great starting points.

Architectural Styles and the Design of Network-based Software Architectures - Roy Thomas Fielding's dissertation, in which the term REST is coined and described. Pretty technical, but a great read on REST and network software architecture in general.

Automatic Multi Language Program Library Generation for REST APIs – An article by Thomas Steiner about the challenge of creating a WSDL-like description language for REST services, including overviews of a few proposed solutions.

How I Explained REST to My Wife – A very straightforward explanation of REST to a layperson, informative and easy to read.

The REST Dialogues and The REST Dialogues, A Real eBay Architect – A conversation between a developer and an imaginary eBay architect, and a response written by a real eBay architect. An easy introduction to the SOAP/REST debate.

Telling the Story behind the Photos

edit

Beth Simmons (CCE/PAL), James Conners (CCE/PAL)

Oceanographic research and data sampling provide the knowledge and means to forecast key ocean-influenced processes and phenomena. These research cruises generate unique but challenging opportunities for education outreach and information managers attempting to convey the excitement of discovery to the general public in near real time. For outreach coordinators it's about unlocking the science creatively and demonstrating how the research benefits current and future generations; getting the public to understand the value of the science behind the research. The added challenges of having limited internet-connection bandwidth to transfer data back to land, as well as time constraints and physical exhaustion all compound the effort. Nevertheless, ocean research cruises provide an opportunity to leverage the innate fascination people seem to have with the open sea with a way to become connected with the science. Pairing outreach coordinators and information managers, helps to maximize the resources available to communicate the science to the public, as well as inspire younger generations to take an interest in the research.

One of the first outreach implementations for Palmer LTER cruises was a project called 'picture-of-the-day' (POD). Its intentions were to bring ship life closer to those at home and advance ocean education through imagery. A photograph during the cruises was sent emailed out with a brief description each day. As the cruises continued year after year, the photographs were collected and made available through a media gallery to both store and to fulfill the numerous requests for images, videos, and content used in other outreach projects. Successful in its design and implementation utilizing limited data-transmission availabilities, this POD project prompted feedback from viewers, connected family, friends and scientists, gave a brief glimpse into what life at sea was like aboard the research vessels and exposed the public to the many oceanographic instrumentation technologies used during scientific research. While appealing, over time there was increased interest to involve the public more readily and build off of the POD concept.

Expanding on the PAL POD experience, the 3rd annual CCE process cruise this year outfitted two teachers with digital cameras, computers and video equipment and asked them to capture life aboard the R/V Melville. Kids, schools and classrooms, families and friends were all invited to interact with the two teachers aboard the ship through a blog where they could ask questions, investigate some of the big ideas surrounding this year's research and also interact with the teachers, scientists and graduate students themselves. Having an identifiable audience and a defined field of interest helped in knowing that we would be able to manage the feedback from participants. Additionally, unlike POD, the blog gave the readers the opportunity to guide the direction of inquiry and help our outreach program understand what the general public was interested in hearing. The teachers posted entries each day and although still nowhere near real-time, the blog replaced the passive reception of news from the ship with an interactive communication forum. What were ordinary daily routines at sea to scientists were now seen through a fresh pair of eyes from a teacher's perspective while aboard the cruise. Furthermore the comments posted by those on land added an additional collaborative point of view. The blogging tool far exceeded our goals, being easy to use and simple to set up. It allowed users to add photographs, short video clips and scientific content. It opened up communication between the ship and shore, built upon existing program components like Picture-of-the-Day and fulfilled the needs for outreach by combining information technology.

Commentary

---

14 Quotes on "Simplicity" by the Greatest Thinkers of All Time.

edit

"The ability to simplify means to eliminate the unnecessary so that the necessary may speak." ~ Hans Hofmann

"The wisdom of life consists in the elimination of nonessentials." ~ Lin Yutang

"There is no greatness where there is no simplicity." ~ Leo Tolstoy

"Simplicity, carried to an an extreme, becomes elegance." ~ Jon Franklin

"The obvious is that which is never seen until someone expresses it simply." ~ Kahlil Gibran

"It is proof of high culture to say the greatest matters in the simplest way." ~ Ralph Waldo Emerson

"The art of art, the glory of expression and the sunshine of the light of letters is simplicity; nothing is better than simplicity." ~ Walt Whitman

"Simplicity is an exact medium between too little and too much." ~ Sir Joshua Reynolds

"Simplicity and naturalness are the truest marks of distintinction." ~ W. Somerset Maugham

"Confusion and clutter are the failure of design, not the attributes of information." ~ Edward R. Tufte

"Simplicity is the ultimate sophistication." ~ Leonardo da Vinci

"Genius is the ability to reduce the complicated to the simple." ~ C.W. Ceram

"Everything should be made as simple as possible, but not simpler." ~ Albert Einstein

"Out of intense complexities intense simplicities emerge." ~ Winston Churchill

Whirlwind Tour of Digital Curation in the UK

edit

Karen Baker (CCE/PAL

Two representatives from CCE and PAL LTER attended the 4th International Digital Curation Conference (DCC) in Edinburgh 1-3 December 2008 as part of the NSF/SES HSD Comparative Interoperability Project. The DCC (www.dcc.ac.uk) is a UK initiative to facilitate data-centric curation by providing an annual international conference, data policy, learning support, curation tools, training and a journal, the International Journal of Digital Curation (IJDC). Conference participants brought diverse perspectives from such fields as institutional data libraries, national data archives, computational software designers, large- and local-scale scientists, social scientists and others. Lynn Yarmey and I traveled to Scotland to present our paper "Data Stewardship: Environmental Data Curation and a Web-of-Repositories" as well as to gain some international perspective on data curation and infrastructure initiatives.

Within the U.S. as well as Europe, there are a variety of contemporary national and institutional efforts with respect to digital data preservation. Most frequently using open-source repository software – Eprints, DSpace, Fedora – national initiatives and institutional libraries are developing digital infrastructure and exchange capabilities with a focus on data preservation. Workshops on the first day of the conference covered the topics of the DCC Curation Lifecycle Model, a Data Audit Framework, and Repository Curation Service Environments. A trio of terms was central to the Lifecycle Workshop and to subsequent DCC discussions:

1. Data Curation – maintaining and adding value to a trusted body of digital information for current and future use
2. Digital Repository – applied to digital storage initiative such as institutional repositories, digital archives, or digital data libraries each with different functionality but having an organizational framework and a technical infrastructure
3. Curation Life Cycle Model – a model that incorporates curation and preservation stages including:
    a. Full lifecycle actions such as description and planning
    b. Sequential actions from creation or receipt through appraisal, ingestion, storage, access/use/reuse and transformation
    c. Occasional actions such as dispose, reappraise and migrate

In our paper, we emphasized the importance conceptually of diverse repositories brought together eventually into a 'web of repositories' but also highlighted the notion of 'local repositories' as a critical but overlooked repository type. With the LTER network model of embedded data management in mind, we established a boundary in order to distinguish 'near' and 'remote' repositories. Repository characteristics were then explored starting with the characteristic difference in terms of a repository's 'distance-from-origin' from the data and the repository's focus, ie on field data or well-defined data objects. Within this framework, the LTER network represents 26 local repositories that are uniquely 'near' in terms of distance-from-origin. Each repository carries out local data curation with a deep involvement in creating datasets for immediate use as well as long-term reuse.

The visit afforded some additional special opportunities. On a walking tour of the narrow winds, we learned about 'gardyloo' and 'hud yer hand' as warning cries to proceed carefully. This brought to mind an early server given the name 'gardyloo'. We were able to meet with those involved with data management at the British Antarctic Survey, a PAL partner, and those at the Natural Environment Research Council (NERC) sponsored British Oceanographic Data Center (BODC). In addition, attendance at a series of talks commemorating the 25th anniversary of the University of Edinburgh Data Library provided further insight into past and present data efforts. Finally, there was a special conference dinner at the Edinburgh castle. With the strains of bagpipes welcoming us, we had a chance to view

some distinctly non-digital artifacts – Scotland's crown jewels – where metadata was provided in the form of a docent's lively narrative that momentarily brought to life a wee bit of Scottish history.

IMs from the file and server rooms, and to the home office and beyond: Reporting our experiences with video-teleconferencing...

edit

IMs from the file and server rooms, and to the home office and beyond: Reporting our experiences with video-teleconferencing and telecommuting.

Linda Powell (FCE), Nicole Kaplan (SGS), James Williams (LNO)

Information scientists have managed information and data over the past several decades on medium ranging from field notebooks to virtual servers, and in file cabinets to service oriented architecture.

Over these years there have been important technological developments, including the local area network, World Wide Web, ubiquitous internet connectivity, and centralized reliable mass storage servers that have changed the way we work and collaborate across the network. Recent network-wide installations of social and collaborative software have created an important opportunity for information managers (IMs) and principal investigators (PIs) in the LTER Network to develop collaborative work environments for our communities. Video-teleconferencing (VTC) has enabled IMs to communicate regularly with other IMs across sites, as well as with IM team members and PIs at their own sites. In addition, such tools have allowed members of LTER sites to remain on staff when administrative offices and home institutions change. Continuity in IMs and PIs is considered an asset and essential to the success of our long-term projects at sites and across the network. VTC and telecommuting have become important tools in maintaining employees, creating flexible work environments, and enabling communication. However, implementation has included both successes and challenges and here we share some stories from sites and across the LTER Network.

An example of an LTER site using collaborative technologies among PIs and staff can be found at the Florida Coastal Everglades (FCE) LTER. The FCE LTER office is located on the campus of Florida International University in Miami, Florida but the sixty-seven senior scientists involved with project are spread across twenty-nine different institutions nationwide. Given that Linda Powell and Mike Rugge, the FCE information and project managers respectively, have been working with distributed FCE LTER personnel for nearly 9 years, it was not difficult for the group to adapt to Linda's relocation from Miami to Tallahassee, Florida. She is now telecommuting from home in Tallahassee and participates in FCE and LTER activities virtually. The use of email and telephone are her major communication tools during FCE LTER scientific collaborations, while VTC adds another dimension to her creative collaborations with the FCE LTER staff on campus.

The success of this telecommuting model is a result of a strong relationship with the FCE LTER Lead PI, Dr. Evelyn Gaiser, and tremendous computer support from Mike Rugge. Dr. Gaiser recognized the importance of having information management consistency for the FCE LTER program and found that Linda's telecommuting gave the program stability in its data management and allowed Linda to continue working in a position she enjoys. Regular communication between Dr. Gaiser and Linda via email and telephone is a key component in this model. A Vonage telephone account has been established where FCE LTER personnel can call a 'local' Miami phone number and the call gets forwarded to the home office in Tallahassee. Linda also travels to the FCE LTER office in Miami on a quarterly basis to meet with FCE LTER scientists and maintain her servers and workstation in the office. Researchers are motivated to schedule meetings with Linda in advance to discuss issues with data.

Telecommuting has proven to be a positive working work model for the FCE LTER and Linda; however there are still some personal and technical challenges with working from home. Working from home requires an extreme amount of discipline and organization. It was important for Linda to establish her office space within the house, designate her 'work' hours, and most importantly get the immediate family to recognize those work hours. There certainly are days where Linda doesn't have any direct interactions with her colleagues, either by telephone or VTC, and there are times when she does miss the 'campus' environment and impromptu meetings with friends and colleagues.

Linda also made major changes in her computer operating systems and invested in Apple's powerful iMac machine for its' large twenty inch screen, portability, built in iSight camera, 2.66 GHz Intel Core 2 Duo processor, and 2 GB 800MHz DDR2 memory standard so she can run many applications simultaneously. Additionally, her home office is equipped with an Apple Time Capsule, a Wi-Fi base station for all the computers in the house and a wireless backup hard drive for the iMac. For backup redundancy, two separate external hard drives are also connected to the iMac. High-speed internet is also extremely important for telecommuting, especially when work involves transporting large amounts of data between home and servers at the FCE LTER office and working across a virtual personal network (VPN). A connection at 5.0 Mb/second has supported data downloads or uploads with the VPN running.

VTC allows Linda to participate with the LTER network information management and NISAC committees on a regular basis. However, The IP address on the home LAN is dynamic, which requires that Linda provide the LTER LNO gatekeeper with the 'new' IP address before she can participate with the LTER network committees' VTC meetings. James Williams (LNO) and Mark Servilla (LNO) provided a tremendous amount of technical support in setting up Linda's home network to handle incoming video calls and with much perseverance, they were able to successfully connect to her system through an open source VTC client residing on the iMac called XMeeting.

The Shortgrass Steppe (SGS) LTER is another example of a site where collaboration software has been critical as a means of communication among scientists. The SGS LTER site was established in 1982, within the first cohort of LTER sites and has always maintained administrative offices at Colorado State University (CSU) in Fort Collins, Colorado. SGS LTER is currently working on a proposal under a new lead PI, Dr. Michael Antolin, while past lead PIs still associated with CSU and SGS LTER have recently moved across the state border to the north to Laramie, Wyoming to work at the University of Wyoming. In addition, a strong partnership exists between CSU researchers and USDA researchers based in Cheyenne, Wyoming, as well as with Dr. Paul Stapp, current PI and active field ecologist who performed his doctoral research on the shortgrass steppe, and now works at California State University, Fullerton. Dr. Antolin was interested in strengthening collaborations with his colleagues and the SGS IM Team recognized the value in VTC technology to maintain productive communication at the SGS-LTER. The SGS LTER schedules weekly science planning meetings in VTC rooms, supported by the Natural Resource Ecology Lab and College of Natural Sciences at CSU, with LifeSize and Polycom units respectively. The CSU units are used for managing VTC with three or less remote participants, while James Williams manages the VTC with the LNO Polycom bridge when more than three participants are in attendance virtually. The weekly meetings to review the SGS LTER organizational structure, develop scientific questions, and plan research studies is an activity essential to the success of the site's operations, science planning, team-building, and proposal writing.

Across the network, members of information management committee (IMC) are using VTC to engage in self-organized working groups to develop best practices, guidelines, and community driven information management products. IMs across the network are now using VTC to discuss issues, complete tasks, and receive updates on progress. In spring 2008, the IMC created a virtual water cooler, where individuals gather virtually about once a month to hear updates from working group leaders, LNO directors, or partners. Discussions around the virtual water cooler keep the IMC up to date on projects and network goals, and provide a venue to exchange ideas and information outside our annual meetings.

VTC is a means for bringing together scientists and educators, was first proposed in the LTER Decadal Plan (LTER Network Office) and encouraged by the LTER Executive Board to reduce travel, enhance coordination, and support progress among standing committees and working groups (LTER Executive Board). Over the past several months, LTER boards, committees and working groups have enacted VTC to perform their duties and the use of this technology has facilitated communication on a more regular basis, which has helped to increase their performance as groups and decrease their travel time and costs. In addition, IMs and PIs have adopted new software tools for collaborating, established more formal and regular, yet virtual practices for communicating, and introduced flexible styles of working, which all have enhanced involvement between site staff and scientists, as well as IMs across the network.

References

LTER Executive Board Meeting Minutes, March 7-9, 2007, Arlington, VA http://intranet.lternet.edu/modules.php?name=UpDownload&req=viewsdownload&sid=58

LTER Network Office Publication #24: The Decadal Plan for LTER, October 2007. http://www.lternet.edu/decadalplan/.

News Bits

PROJECTDB – Planning and Development of a Collaborative Programming Effort

edit

Jonathan Walsh (BES), Jason Downing (BNZ)

The ProjectDB is a means of tracking and cataloging the projects being undertaken by Long Term Ecological Research Network sites. Information managers may recognize it as similar to previous “db” projects such as SITEDB wherein information is agglomerated and made available as a whole. This makes it possible to draw comparisons and similarities to the sites as well as gather data from all the sites to represent the network as a whole.

Although PROJECTDB is similar to SITEDB, STREAMDB and CLIMDB, it’s development is being done with a notable difference. PROJECTDB is being written collaboratively. It will be designed and programmed as a group effort among numerous LTER sites. This brings forth the benefit of diverse knowledge and expertise while including accommodations for individual site differences.

Goals and Scope

The project will create a common framework to express, track, analyze and report research projects at LTER sites. It will use XML as the mechanism for exchanging information.

The resulting system will consist of web clients for each site in popular languages such as Java, ASP, PERL, etc., which will be able to retrieve the project information. The information will be transformed in two ways – one to transform it from XML to HTML and the other using individual style sheets to produce a human-readable web display with site specific layout and color scheme.

A harvester, just like in CLIMDB and STREAMDB will be constructed for sites that wish to use their own legacy database engines to store their project data. These sites will produce XML files for their projects to be harvested. It will rely on the EML schema and check for well-formed structure upon intake.

Information managers were polled to see if they would be interested in being involved in such a project and if so, which component or components of the project. The project phasing was as follows:

1. Define functionality
2. Implement and test
3. Provide usability feedback
4. Write code

Two workshops were planned. The first was to develop use case scenarios, define methods, and identify necessary changes to the Ecological Metadata Language (EML) schema. That workshop was held in November.

The second workshop will consist entirely of programming.

Technology for this project

This project will manage a database of LTER research project information. Ideally that will consist of structured project descriptions including but not limited to:

- Description
- Datasets/Metadata
- Products
- Goals
- Rich text description
- Photographs
- Links to Investigators’ pages

<img src="http://databits.lternet.edu/files/image001.gif" style="height: 309px; width: 250px;">

eXist is a database system written entirely in Extensible Markup Language (XML) and was favored for use with this project. Web Services will be employed in order to provide transport from the database (or perhaps more accurately, databases) of project information to the clients. Utilizing web services will allow a great deal of flexibility and accommodate various means of data feed, thus supporting legacy systems.

Web Services: REST vs. SOAP.

There are two major Web Service transport schemes being used: Simple Object Access Protocol (SOAP) and Representational State Transfer (REST). Briefly, web services are applications that can be hosted on a remote machine and executed remotely, returning some result. The sign-in for Microsoft Hotmail, for example, is a web service. REST was determined to be more appropriate because of its reduced dependence on external sources and more straightforward integration with http.

Advice from Sven Bohm was to use REST. “REST is just basically http requests,” said Sven, “SOAP maps more closely to a procedural standpoint. REST maps more to everything being an endpoint. In REST, there are four basic verbs – PUT, POST, DELETE, and GET. So SOAP has few nouns and many verbs and REST has few verbs and many nouns.”

The REST interface to search the database was determined to be best represented as such:

Basic search:

- Search creator, keywords, title, abstract;
- Full text search (possibly slow);
- Support for Boolean operations between words.

Advanced search:

- Creator
- Keyword (Use controlled vocabulary for each different keyword set)
- Title
- Abstract
- Temporal, e.g. year
- Spatial
- Funding agency
- Organization
- Support Boolean operations between those search fields

The REST search service should then return a browse list: ID, creator, title, keywords, temporal coverage, and alternately the full document.

Database selection: eXist

<img src="http://databits.lternet.edu/files/image002.jpg">

eXist is a database management system written entirely using XML. This database works easily under the Apache TomCat web server.

Corinna Gries, who is familiar with eXist says the internal user management is a bit primitive but other than that it will do the job nicely. She currently keeps the data list for Central Arizona Phoenix (CAP) on it with over 700 EML documents. Keyword searches are fast using a JAVA interface.

Demonstrations of eXist are online at: http://exist-db.org/

Metacat was also discussed as an alternative database but after discussion with Inigo San Gil, Mark Servilla, and Corinna the group felt eXist was a more appropriate system but admittedly it could be done with either.

AJAX for data entry and presentation

Wade Sheldon presented the pros and cons of Ajax (Asynchronous JavaScript and XML). According to Wade, it's a very efficient way to "skin" a website. By creating an XML style sheet (XSLT) the interface can be made to look easily like the calling page. So each LTER site can easily customize the PROJECTDB interface to look like their existing website. Additionally, the style sheets could be "manufactured" and given to the LTER site webmaster who would then only need to know how to deal with XML generated on their own system.

Online EML Editor

Central Arizona Phoenix and the Network Office are jointly working on an online EML editor. Information stored – in EML format - on the eXist database can be maintained with this editor. The editor will provide an online interface and forms to simplify the process of managing this information. This editor will provide a way to access the information in the PROJECTDB eXist database but LTER sites that choose to keep their legacy data in their own database format can provide their own means of editing and still participate in PROJECTDB.

Use Cases

Use case scenarios were developed to determine the requirements for the various elements of the system.

Ken Ramsey researched the writing of Use Case Scenarios and found a good guide online. The URI for the guide, "Writing Effective Use case Examples" is here: http://www.gatherspace.com/static/use_case_example.html

The group worked together using video teleconferencing and developed use cases for the system. It turned out to be an iterative process in which actors, goals, and other attributes that make up a use case scenario were identified.

In one example, the use case scenario in which a research project is described, the following key elements can be identified:

Title: Describe a Research Project
Goal: Provide the ability to collect and organize information for annual reports using information about active research projects.
Actors: site manager, site leadership, report writer, administrator.

Once identified, the use cases can be placed into a grid and can then be ranked by such factors as scope, complexity, and priority.

As the use cases are identified and the elements are attributed to them, there becomes a framework that can be passed to a software team to most effectively go about producing an application that will meet all the needs. This is in progress at the time of this writing. Ken Ramsey, Suzanne Remillard and Kristin Vanderbilt are the lead authors.

Specifications

The generation of specifications will follow the completion of the use case scenarios. The team is using the use case structure and methods set forth by Ramsey and using Google Docs to collaboratively edit the document. Another collaborative tool being used in this effort is the use of a version control system (See SVN, below)

Possible change to eml schema

It was determined that the existing EML schema will need modification when it comes to describing projects. Margaret O'Brien is working on changes to the EML schema that will help facilitate the PROJECTDB system. One example is the EML Project descriptor: In the Project description node there is an attribute titled 'KeywordSet' that will be given an optional NAME sub-attribute to allow tracking of the Habitat/Ecosystem/System for the project.

SVN

Subversion (SVN) is being used to help coordinate with multiple programmers and multiple revisions. Subversion (as opposed to attempts at overthrowing authority!) is software for revision control. This software is being fostered by Mark Servilla of the Network Office for this project. It allows tracking of all work and all changes to all elements --- even if done in different languages. It allows the project to be "rolled back" to any point in time. It also allows developers to be apart from each other in space or time as they work together on a common goal. Subversion is being adopted by the Network Office as a replacement to the older standard revision tracking system, Concurrent Versions System (CVS). Subversion has a more advanced storage format, and is faster.

Project Workshop Afterthoughts

This type of network collaboration approach has proven to be an effective technique in developing information management tools and continually shows more promise for the future. That being said, this project would not be possible without the support of the LTER Network Office, James Brunt, and the entire LNO staff. The use of the LNO facility enables project participants to focus efforts on specific issues for a dedicated time period while in supportive environment with ample resources (technological and human). The Polycom VTC system allowed for participants who were unable to travel during the workshop dates to have meaningful interaction and make significant contributions to the team products. Special thanks to James Williams for his efforts to keep the remote articipants connected.

Good Tools And Programs

Creating Archival Documents

edit

John Porter (VCR)

When is a WORD .doc file not a WORD .doc file? This was a question that confronted me when I attempted to open up an old copy of a letter I'd written in 1994. The letter was written using Microsoft Word, complete with the .doc extension, but my new Microsoft Word 2007 refused to open it. After drilling down through the Microsoft help I found the following statement:

> *You try to open a file that was saved in one of the following earlier Word formats. Word 2007 no longer supports documents that were saved in the following Word formats:*
> *• Microsoft Word for Windows 1.x*
> *• Microsoft Word for Windows 2.x*
> *• Microsoft Word for the Macintosh 4.x*
> *• Microsoft Word for the Macintosh 5.x*

Fortunately, OpenOffice Writer was still able to read the letter in question, so I was able to proceed, but this bought home to me, once again, the difficulties associated with maintaining digital information over time. After all, if you can't count on a market dominant software package to maintain backwards compatibility for at least 20 years, who can you count on? With that in mind, here are some ways that I try to assure archival readability of documents.

Best practices dictate that a document of potential archival value be stored in a form that will only require infrequent updates (if any) and that, if all else fails, information can still be extracted from the document. For word processing documents probably the best choice of format is the Rich-Text-Format (.RTF). This is an ASCII text file that, in addition to the text, contains commands for formatting such as \par to define a paragraph. It is widely read by a variety of word processors, but because it is ASCII it can be read using even the simplest editor, and text can be read or extracted even if all the text formatting settings are ignored. Hypertext markup Language is another alternative, but is less flexible in being able to handle complex text formatting, and stores graphics externally, rather than as an encoded part of the document itself.

Spreadsheets are particularly difficult to successfully archive. Although comma-separated-value (.CSV) files can preserve data in a text form, any built-in formulae are difficult to retain. In the past, the best approach was to store each spreadsheet in as many forms as possible, hoping that at least one would be readable at some future time. However, the advent of the OpenDocument Format (ODF) is providing at least some hope that a single format may be suitable. ODF documents (which can be wordprocessing, as well as spreadsheets) consist of a compressed "zip" file containing eXtensible Markup Language (XML) files and binary image data. The new Microsoft Office 2007 uses ODF-like files, conforming to a "Office Open XML" (OOXML or OpenXML) standard that also include XML and image data, but do not use ODF schema. Among the differences, the Microsoft .docx, .xlsx and .pptx formats use XML tags with extremely short names (to help with speed, at the cost of interpretability). Fortunately, there are increasing numbers of translators that can interconvert between these documents. Accessing the XML files for both these formats requires a program that can "unzip" the container files. The "Zip" compression scheme was created in the mid-1980s and has remained remarkably robust – in part because it began life as an open standard. Its adoption into current standards means that its life is likely to be extended. However should it ever "disappear" we are going to lose access to a lot of data!

Where editability is not critical, Portable Document Format (PDF) files are a popular choice for archiving document data. Starting as a proprietary format from Adobe Inc., it has now become an ISO standard. The widespread use of PDF documents and a wide availability of tools to create and read them means that, at least for the next 10 years or so, PDF documents will remain a good archival format. Beyond that timeframe things are less certain, but the widespread use of PDF documents guarantees that there will be several options for moving today's PDFs to tomorrow's standards. There have been several versions of PDF standards, but the earlier versions remain readable by the most recent viewers. To some degree this is because of the use of the Postscript text formatting language as the basis for PDF documents.

For relational databases, my favorite way of archiving data is to use a "dump" command to create a text file containing all the SQL commands needed to recreate the entire database. A second alternative is to export tables into ".DBF" files. DBF files made their start in life as the underlying data structure for the commercial DBASE program. However, they are now widely used in many different applications (the attribute tables in ArcGIS Shapefiles are .DBF files), where a light-weight database is needed to capture both structure and content of data.

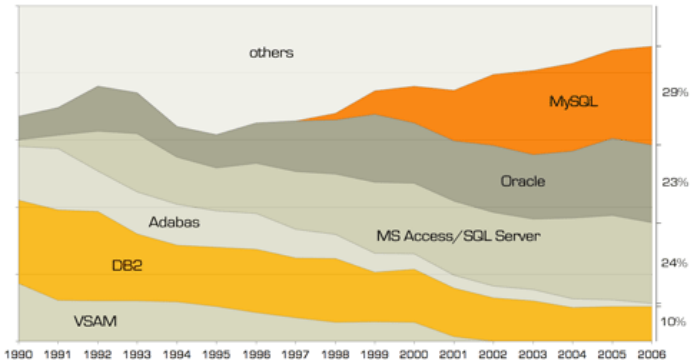
Ultimately, continued curation by knowledgeable information managers is the best assurance of archival readability of data and documents. However, hopefully the tips above will decrease the frequency with which archives need to be transformed from one form to another.

MySQL Workbench: A Visual database design tool

edit

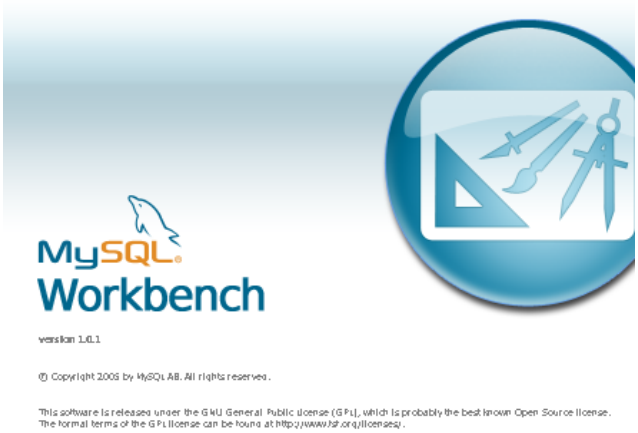
James Conners (CCE / PAL)

Relational databases make ideal data storage backends for many applications. MySQL in particular, being free and being supported by many programming languages, is especially popular. And being so popular



Source: JoinVision E-Services GmbH, July 2006

there has been much interest in developing applications that support working with MySQL databases. In addition to the command line interface built in to the software, there are many available programs both free and commercial for creating, updating and managing MySQL databases. And so as a result it is often times the case that within a team of developers each one has his or her preferred interface. Certain tools seem to be better suited for different tasks. A consequence, however, is that many times updates made throughout a particular database's life cycle are done using different software, sometimes introducing inconsistencies. For smaller applications that require rapid development and deployment, this may not be a problem. When the scale of the project increases along with the number of developers maintaining the database(s), it often becomes more important to establish best practices along with some way of reviewing changes and the state of the project.



One application I've recently found useful for creating and maintaining larger-scaled database projects is MySQL Workbench, a visual design tool and successor to DBDesigner4. Specifically, we've recently updated one of our largest databases containing almost forty tables using this software. With the free version the capability is available to reverse engineer an existing MySQL database using the CREATE script that you can export using almost any database management application. Using this script the software will build a model, including an entity-relationship (ER) diagram, for the database. Existing tables can be edited and new tables added through the visual interface. The database can then be updated by exporting an ALTER script. The commercial version has the capability to connect directly to an existing database so that changes to its structure can be made directly through the visual interface.

Editing a database's schema through the ER diagram interface allows a perspective that working through other interfaces doesn't. Also, schema components can be grouped visually and organized in a way that provides for more comprehension by viewers and facilitates design discussions among developers. Note, exporting schemas can be done as images or PDFs, adding a valuable component to a project's documentation.

Another one of this software's more appealing features is the ability to write scripts and plugins for task automation and extending its capabilities as an application. It currently supports scripts written in Lua (http://www.lua.org/), but support is being worked on for other languages such as Python. Being able to augment the software by writing custom scripts improves its ability to support changes in development standards and practices over time.

We've yet to explore the majority of the features of this application or incorporate its use into our common development practices, but there does seem to be a lot of potential there. If nothing else, it's a great tool for quickly creating ER diagrams for existing MySQL databases, for sharing with collaborators or to for documentation.

Links and references:

- MySQL Workbench (Developer site): http://dev.mysql.com/workbench/
- Visual modeling article: http://adtmag.com/article.aspx?id=10759&page=
- Lua programming language: http://www.lua.org/

Mapping your mind with FreeMind

edit

John Porter (VCR)

Tired of the same old outlines? “Mind Mapping” is a recent approach to developing outline-like documents and figures that exploit the organizing capacity of hierarchies, but eliminates the linearity of traditional outlines, where jumping from one outline heading to another may require paging down. Mind mapping programs typically present a hierarchy as an “octopus” whose legs become successively divided as you proceed down the hierarchy. For example the outline:

VCR LTER Research Sites

- *Islands*
  - *Hog*
  - *Smith*
  - *Metompkin*
- *Mainland*
  - *Marshes*
    - *Phillips Creek*
    - *Cobb Mill Creek*
    - *Fowling Point*
  - *Upland*
    - *Eastern Shore NWR*
    - *Brownsville Farm*
- *Lagoon*
  - *South Bay Seagrass*
  - *Hog Island Bay*
  - *Gargathy Bay*

Would be represented as a “Mind Map” as shown below:

Advantages are that there is less priority among elements at the same level and your “view” is able to take in more at one time. You can also toggle on or off the children of any branch to provide a clear, expanded view of what you are focusing on. It is less suitable where outline entries are longer (e.g., full sentences), but these can be accommodated, even if they make for a less attractive map. Moreover, entries can be linked to external resources using hyperlinks or highlighted using added icons.

There are several commercial (e.g., MindManager) and open source (e.g., VYM) mind mappers. A favorite of mine is FreeMind, which is an open-source, JAVA-based mind mapper that is available from: http://freemind.sourceforge.net . It features a friendly interface, and numerous formatting and export capabilities (the outline above was actually exported directly from FreeMind into WORD as a copy/paste operation). Its underlying data structure is XML-based, so it is relatively easy to automate the writing of mind maps. For example I used it to create a browseable FLASH map of our data sets (http://www.vcrlter.virginia.edu/data/datafigs/topicmapflash.html).

There are also a large number of third-party tools that work with FreeMind for importing text outlines, working with bug trackers, creating CSV files etc. at: http://freemind.sourceforge.net/wiki/index.php/Import_and_export_to_othe... .
I’ve enjoyed using FreeMind as a way to “brainstorm” problems in a way that can be easily exported or even posted as an interactive web page, and its flexibility and ease of use makes me think I'll be using it quite a bit more in the future!

Good Reads

Enabling long-term oceanographic research: Changing data practices, information management strategies and informatics

edit

Nicole Kaplan (SGS)

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

Authors Baker and Chandler have helped to shape the components of information systems within long-term oceanographic research projects and served as data stewards for the Long Term Ecological Research (LTER) Network and US Joint Global Ocean Flux Study, respectively. In this article, they create case studies by discussing their work with multiple institutions and ocean and information scientists, and highlight strategies to building successful information ecosystems.

Professionals in their positions disseminate data and information at local and global levels within a world where data are expected to be available for reuse, and there is not always a clear technological answer to provide access and support integration. They illustrate that information systems are complex and interact with human, as well as environmental systems.

These case studies may help others in similar roles to describe components of good data management practices, identify challenges related to reuse of data by a wider and remote audience, and enhance local understanding and communications around information management strategies. The examples from long-term projects include the recognition that it is important to incorporate information management in science planning to facilitate data access and data integration beyond the local study and into the future. Information management systems, such as the ones described in this paper, have grown from providing data organization at the site to manage information exchange electronically. Baker and Chandler explain the importance of communicating local knowledge about the data to support efficient and effective data reuse. They also report that local data and metadata are now accessible in repositories that contain information from beyond the local site. They discuss the importance of information awareness, flow and federation as a site or project’s information is included in larger and broader archives and repositories.

This paper describes different and similar approaches to information management from two long-term projects both focused on ocean science and both part of institutions participating in multi-site interdisciplinary research initiatives. After gaining much experience in developing their own systems, policies, infrastructure, and working environments, Baker and Chandler describe the complexities inherent to managing diverse information for local and global users, within their own systems and across a federated information landscape. They illustrate the concept of an information environment where ocean domain and information scientists can discuss elements comprising their systems; some are data, technology, oceanography, and people, and the ocean informatics that occur where these components interact.

Disputed Definitions

edit

Karen Baker (PAL/CCE)

This short news feature in Nature (2008, Oct 23. Disputed Definitions. Nature 455: 1023-1028) presents five example terms commonly used in the scientific realm. The examples - paradigm shift, epigenetic, complexity, race, and tipping point - are a mix of discipline-specific and interdisciplinary terms. Each is defined briefly by one of the co-authors and shown to have a number of interpretations. For instance, in the case of ‘tipping point’, the term is explained as being used with two very different meanings:

1. An irreversible point of no return
2. A threshold at which a reproductive rate goes above one

For the terms, the article introduces two types of definitions, both susceptible to ambiguity: stipulative or working definitions and essential definitions that identify uniqueness or characterize difference. The ambiguous terms are presented as exceptions since there are a plenitude of terms that are not so ‘troublesome’.

The article is a quick read and an interesting one, especially with semantics playing an increasingly important role for all those working with data whether addressing metadata in particular or ontological issues in general. The article seems to miss, however, the opportunity to make explicit the idea that ambiguous terms may be identified as areas of active knowledge building in contrast to discussing them as problematic. The article sets the stage from the start with a title

that identifies definition differences as 'disputes' perhaps as a journalistic device for drawing in the reader. An alternative title such as 'Ambiguous Terms Represent Knowledge-Making Arenas' would have foregrounded knowledge making processes. This change in focus transforms the 'disputes' into 'inquiries' and 'healthy tensions' that are integral to scientific research. While the semantic work associated with data and metadata has exploded onto the scene in recent years, the work of knowledge-making remains largely implicit as is the case in this article. That is, in addition to highlighting a lack of consensus on contemporary topics, it is also valuable to focus on articulating what is an ongoing, exciting arena of scientific inquiry. From this perspective, ambiguous terms are placeholder concepts, indicators of arenas where scientists are participating in integrative work and where time is required for describing, classifying, and synthesizing the intellectual landscape. Explicitness may well shape how participation in semantic issues unfolds, though wishful thinkers may hope this knowledge work will be done "elsewhere" - perhaps as a comprehensive automated mapping or maybe an agreement negotiated by a disciplinary 'semantic disambiguation' working group of some type. Though lacking in semantic scope, the article 'Disputed Definitions' alerts us to the existence of terms with multiple meanings and prompts us to consider the ramifications of ambiguity in scientific definitions.

Digital Data Practices and the Long Term Ecological Research Program Growing Global

edit

Anne Graham (Civil & Environmental Engineering Librarian, GIS Liaison to the Engineering & Science Libraries, MIT)

Data curation is "a broad, integrative framework" and "a complex role."

Karasti, H. and Baker, K.S. "Digital Data Practices and the Long Term Ecological Research Program Growing Global." International Journal of Digital Curation. Vol. 3, No.2, (2008)
http://www.ijdc.net/ijdc/article/view/86/104

As data collection, management and preservation evolve, so do those who provide access to data. This article tells the story of this evolution within the Long Term Ecological Research (LTER) community. The history of the LTER program dates back to 1980, when the National Science Foundation originally provided funding for long term ecological research at sites within the United States. In the early 1990s it became an international effort. Because of its history and experience the LTER community has a lot to teach the data curation community.

The article describes the detailed, time intensive, critical work that information managers in the LTER provide. It identifies the domain differences between ecology, with its set of wide databases (smaller datasets of diverse data) and genetics, with its deep databases (huge datasets but more uniform data). It talks the reader through the process of working with local data and preparing it for global access.

Karasti and Baker describe the specifics of curating ecology data while providing relevant and thorough knowledge to other disciplines/domains that create and mange data. They explain the work of LTER information managers through insightful quotes, and create a story that is at once overwhelming and well defined. By providing details they demystify a multi-scale process, and deliver a complete picture of the state of their operations. And they supply wisdom that can only come from three decades of "synthesis endeavors[, which] have brought to the fore what may be described as semantic and sociotechnical issues relating to exchange, integration, and interoperability."

A three part role is portrayed, which includes "data care, science partnering and information infrastructure work." As the scope of the program expands (from local to global), so do the roles performed by information mangers, with rapidly evolving technology opportunities creating new expectations. The balance needed to grapple with the accumulated scope and proliferating opportunities will require a lot of communication between many people with different skill sets, all working together toward the ultimate goal of providing data to present and future scientists.

Calendar

Events

edit

Event: International Conference on Ecological Informatics

Location: Cancun, Mexico

Dates: December 2-5, 2008

Web: https://conference.ecoinformatics.org/index.php/isei/isei6/

Event: Social Ecological Systems Workshop

Location: San Juan, Puerto Rico

Dates: December 16-18, 2008

Event: Hawaii International Conference on System Sciences

Location: Hawaii

Dates: January 5-8, 2009

Web: http://www.hicss.hawaii.edu

Event: LTER Science Council meeting

Location: San Diego, CA, USA

Dates: May 13-14, 2009

Web: http://intranet.lternet.edu/meetings/

Event: Scientific and Statistical Database Management

Location: New Orleans, LA, USA

Dates: June 2-4, 2009

Web: http://www.ssdbm.org

Event: Ecological Society of America Meeting

Location: Albuquerque, NM, USA

Dates: August 2-7, 2009

Web: http://www.esa.org/albuquerque

Event: LTER Information Management Committee Meeting

Location: Estes Park, CO

Dates: September 13, 2009

Web: http://intranet.lternet.edu/im/

Event: LTER All Scientists Meeting

Location: Estes Park, CO

Dates: September 14-16, 2009

Web: http://lternet.edu/asm

Event: American Society for Information Science and Technology

Location: Vancouver, Canada

Dates: November 6-11, 2009

Web: http://www.asis.org/conferences.html

Event: Digital Curation Conference

Location: London, England

Dates: December 2009

Web: http://www.dcc.ac.uk/events