01001100 01010100 01000101 01010010

LTER DataBits
Information Management Newsletter of
The Long Term Ecological Research Network

01001100 01010100 01000101 01010010

◆ **Feature Articles**

About this Issue

NEON is Bursting Out All Over!

Web Services Workshop (February 2-5, 2004)

Postnuke Portal Software: Community, Content, and Collaborative Management System

◆ **Good Reads**

Data Grids, Collections, and Grid Bricks

The Dry and the Wet

An International Framework to Promote Access to Data

DataBits: An electronic newsletter for Information Managers ----- Spring 2004 Issue
(http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/04spring/)

Featured in this issue:

John Porter talks about the latest regarding the National Environmental Observatory Network (NEON) program. Jonathan Walsh shares his experience at the Web Services Workshop (February 2-5, 2004). Karen Baker, Shaun Haber and Marshall White discuss the Postnuke Portal Software.

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 to incorporate rotating co-editorship. Availability is through web browsing as well as hardcopy output. LTER mail list IMplus will receive DataBits publication notification. Others may subscribe by sending email to majordomo@lternet.edu with two lines "subscribe databits" and "end" as the message body. To communicate suggestions, articles, and/or interest in co-editing, send email to databits-ed@lternet.edu.

----- Co-editors: Chi Yang (MCM), Jonathan Walsh (BES)

◆ Feature Articles

NEON is Bursting Out All Over!

- John Porter (BES)

After a quiescent phase caused by uncertainty about the future of the program, the National Environmental Observatory Network (NEON) program is showing new life - with new implications for Information Managers. For those unfamiliar with NEON, its aim is to provide the infrastructure needed to help address the nation's environmental challenges through research. It has been proposed by NSF for funding under the Major Research Equipment and Construction program for the last several years, but has yet to receive funding by Congress. If ultimately successful, it will provide hundreds of millions of dollars to establish ecological observatories.

Several recent developments have sparked new interest in NEON. The first is a National Academy of Sciences report (http://www.nap.edu/catalog/10807.html) that strongly supported the concept of NEON. Second are the activities of the Infrastructure for Biology at Regional to Continental Scales (IBRCS) project, of the American Institute of Biological Sciences (AIBS) (http://www.aibs.org/ibrcs/), who have developed white papers and conducted public meetings that have refined the NEON concepts. Finally, there is the request for proposals issued by the National Science Foundation for a $6M, 2-year intensive planning grant that is expected to be funded later this summer.

Why is this of special concern to information managers at field stations and LTER sites? One reason is that NEON will put unprecedented pressure on our ability to manage information, both by increasing the quantity and quality of data and by demanding that data be more accessible by researchers. For this reason the IBRCS white paper (http://ibrcs.aibs.org/reports/pdf/NEONCoordRpt.pdf) calls for between 25 and 75% of NEON operation budgets to be devoted to information management. Potentially this could be over $20 million per year!

A second reason is that increasingly regional groups interested in NEON are self-organizing observatories, with the objective of being ready to play a role in NEON when funding becomes available (http://ibrcs.aibs.org/neon/regional-index.asp). Many of these groups (this spring, I've personally attended meetings for the Mid-Atlantic Region Ecological Observatory, the South East Ecological Observatory Network and the Neotropical National Ecological Observatory Network), have been holding workshops and meetings. Almost invariably, a primary topic has been information management.

NEON will provide both new challenges and new opportunities. As a group, we need to think aggressively about how ecological information management can be moved to the "next level" if major new resources become available. If we aren't prepared, then it is likely that we won't play a role, and that would be bad both for the ecological community and us!

Web Services Workshop (February 2-5, 2004)

- Jonathan Walsh (BES)

Introduction

The Long-Term Ecological Research Network has been working with San Diego Supercomputer Center on several initiatives designed to enhance our data management and distribution capabilities. Under one such initiative, a group of LTER information managers assembled at the San Diego Supercomputer Center for a two-day training workshop. The purposes of the workshop were to give hands-on experience to the participants and to discuss larger implications of the technology such as the "grid" concept (explained in the Spring, 2003 edition of DataBits at http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/03spring/) of information sharing. Additionally, each participant was to leave with a functional web services application on their laptop.

Web Services

Web services are modules of code living on machines hooked to the web that listen for requests. We use them as little appliances so we don't have to rewrite code over and over; we can all use the same code. There are many potential applications for web services within the LTER network, and between the LTER network and the wider scientific and management communities.

The majority of the Web today is dished up by a simple markup language. Two actual machine commands, GET and POST pretty much control the whole thing. And each web offering is essentially a separate endeavor. The problem with programming is that it gets complicated and hard to read. It's made easier by breaking common tasks up into reusable blocks, or modules. Web services are reusable building blocks. They are "listeners" waiting for a request for service and are self-contained and self-describing. They can be combined and they can use each other, reducing and in many cases eliminating the need to write new code. This can dramatically increase efficiency and distribution. The concept of web services wasn't very interesting a few years ago. But now that bandwidth and disk storage are cheap it is becoming an increasingly attractive means of sharing information.

Microsoft's "Passport" is an example of a web service. A variety of sites, including "Hotmail" use the Passport service to handle username and password authentication. Sites using the Passport web service do not have to waste any time programming the interface and the mechanism of password entry. Another example: imagine there is an established module, let's say for sharing snow reports. Machines, specially geared to offer snow reports would be sitting around the world listening for requests for their snow report information. Websites could display the snow report information without having to know anything about how to actually compile such information. So your favorite travel website could offer snow reports at all the world's ski areas without ever having to do the work of measuring snow.

The Workshop

The implementation consisted of breaking into three groups - climate and hydrological data, land-use mapping and lakes. The groups first discussed application to existing systems, needs, and potential web services to answer specific scientific questions. Each participant had prepared their laptop for web services prior to the workshop. This involved installing server software, a Java development toolkit, drivers, and adjustment of the machine's "environmental variables". Then each participant developed their own web service, accessed their information remotely over the web and displayed it.

During the course of the two-day workshop, a variety of issues, opportunities, and gaps were identified and discussed. These are presented below in the following categories: potential applications to existing

systems, needs, issues concerning the use of web services, potential web services specifically geared toward scientific questions

Potential applications to existing systems: What are the possible ways web services could be incorporated into a site's information management system?

- Harvesting of data from distributed sets
- Accessing a central database (note: it was also determined that a real advantage of web services is to allow the use of distributed data instead of a centralized database, yet a web services module makes sense for a centralized database as well.)
- Aggregation services - smoothing, monthly averages, min/max?
- Services to get metadata
- Resolution of individual observations, lists of parameters
- Spatial coverage selection
- Links to external data, such as different disparate sources for precipitation data.
- Quality assurance and control - Inter-site comparisons, filling gaps in datasets, neighboring stations, regressions?
- Query and visualization (e.g. based on characteristics of site, elevation of site?)
- Calculation services - general statistics, analytic procedures
- Visualization

Needs: What do information management systems need to consider before implementing web services and what are the broad LTER information management needs that can be addressed by web services?

- General query tools for data and metadata
- Metadata at site, station and sensor levels
- Web services specifics - for harvest, query...
    - Parameters
    - Methods
    - What is returned (ASCII, XML, record, file)
    - Web services standards
    - Quality assurance
    - Feasibility of all sites running web services

Issues concerning the use of web services: What problems might arise form the adoption of the web services technology?

- Web services standards
- How is data currently stored
- Feasibility of all LTER sites running web services
- Quality assurance

Potential web services to answer scientific questions: What scientific research questions could be answered with the use of web services?

- higher level summary data/statistics
- spatial and temporal analysis
- GIS Linkages to databases
- Cross site comparisons

- Identification of peak points, such as peak discharges during storm periods

Lessons Learned

It turned out that most of the computers were not able to function as web services servers without a good deal of additional tweaking. This was mainly because there are many inherent complexities that cannot be identified until actually trying it. Eventually they all worked. Most Information Managers will not have to actually write web service applications, and it is likely that the technical problems we faced will be resolved prior to broad use. It can be reasoned that a realistic goal for the Long Term Ecological Research Network would be to simply compose a list of web services that would be desirable. One such web service might be a set of general query tools for data and metadata.

There is no doubt, however, that web service technology is highly useful and eventually we will all use web services in whatever form they take. For example, PostNuke, a weblog/Content Management System (CMS) uses the simplest form of web services - Really Simple Syndication (RSS). RSS allows content from many sources to be gathered onto a single website. The currency of the information is controlled by the sources. Note, however, that the accuracy of the information is no longer in the hands of the website displaying it.

There are barriers to the adoption of this technology within the LTER network. It can be tricky to set up, it requires sources of good, accessible data and metadata, and it can take a considerable investment of time on the part of the Information Manager. On the other hand, there are potentially large payoffs to adopting the technology. So while there is little doubt web services will become widely used, adoption by LTER research sites will most likely take some time.

Postnuke Portal Software: Community, Content, and Collaborative Management System

- Karen S. Baker (PAL), Shaun R. Haber (PAL) and Marshall White (NET)

Introduction

Postnuke, a web software application toolkit, is a community, content, and collaborative management system (C3MS). A Content Management System (CMS) is an efficient tool for setting up and managing the content of a website. Simple usage may include adding a static web page or updating a calendar; an advanced application may involve maintaining a complex network of community-based forums.

Postnuke provides extensive tools for generating dynamic web pages and enables a community of users to effectively collaborate on posting web content. With a simple user interface and straightforward navigation, Postnuke creates project portals. It provides viewers with convenience, security, and stability for retrieving and posting data.

Architecture and Modules

Postnuke, (http://postnuke.com) an Open Source Project written in PHP, is a server-side scripting language that is integrated with Apache web server and MySQL database. This software suite may be launched within a Linux, UNIX, or Windows environment. The software is released under the Gnu General Public License (GPL); it is free to download and alter. All pages generated with Postnuke can be configured via a web browser, allowing web managers to work remotely on a site. Further, any user with a

registered account may post content in sections of the site, depending on the established permissions.

Postnuke emerged as a program fork from PHPNuke, which itself was released in June 2000. It is under active development as users around the world contribute modules to the developing community. An Application Programming Interface (API) exists which means users have access to system capabilities for design of contributed modules.

Postnuke has three primary components: *module*s, *blocks* and *themes*. The *modules* are applications that allow users to do something. Postnuke modules include:

- Calendar - displays previous and upcoming events
- Content Express - adds static content (docs, minutes, etc.)
- Discussion Board - provides user mechanism for continuous forum-based discussions
- Gallery - organizes collections of photos
- Login - provides user registration and login accounts
- RSS/Syndicate - syndicates site content as an RSS feed

Blocks allow information to be presented in various locations (i.e. left, right, and center). *Themes* control how the pages look and feel. In combination with the MySQL database backend, there is a separation of content from presentation and business logic.

The growing list of Postnuke modules provides functionality and interactive layers for a website. Some modules may create simple blocks or sections on a webpage while other modules may present an entire new page for the site. A few examples of modules are summarized in the box. Additional functionalities include forums, mailing lists, searches and survey capabilities.

Emergent Software Capabilities and Use
Postnuke's growing user community includes use at several LTER sites. John Porter (VCR) is an early prototyper who makes use of the user upload capabilities and the project description aspect (see John's site at http://www.mareo.org). Marshall White (NET) who has provided design and implementation support over the years to multiple sites, sees the package as a potential method to provide a 'web-in-a-box': "One of the goals of using Postnuke is to ease the burden on the data manager at a site". Kristen Vanderbilt (SEV), in collaboration with the Network Office staff, created a Postnuke web site (http://sev.lternet.edu) incorporating the calendar module. Nicole Kaplan (SGS), Ken Ramsey (JRN), and Kristin Vanderbilt (SEV) have discussed cross-site efforts while Palmer site is prototyping the package because of the availability of collaborative software tools such as blogs and wikkis that enable community participation.

User experience leads to the to the suggestion that a mixed approach to web sites be taken, making use of CMS in conjunction with other web tools. In addition, the RSS module provides a potentially powerful mechanism to share information across sites using a simple webservice. As new information gets posted to a site, the information can automatically trigger generation of an XML news feed that can be aggregated at other sites or the LTER home page. This provides a new distribution mechanism for LTER community information that can be explored and prototyped.

Portal management is an emerging technology. A number of C3MS type software options exist or are under development with features and functionality similar to that of Postnuke. For instance, there is EZPublish with an object like framework and Zope which is Python-based. Additional web management

systems include PHPNuke and Drupal. Drupal is both modular and has design templates for ease of administration via the web. In the planning is a more XML based approach by the Apache Software Foundation that is in the process of planning to vote on a CM package.

With web site management packages today, there is the typical quandary of choosing from a group of options and then working with a product under active development. The open source community constitution of Postnuke combines with its ease of installation and toolkit approach to make it an interesting candidate to consider.

Good Read: Data Grids, Collections, and Grid Bricks

- Karen Baker and Jerry Wanetick (PAL)

Rajasekar, A., M. Wan, R. Moore, G. Kremenek, and T. Guptil, 2001. Data Grids, Collections, and Grid Bricks. in 20th IEEE/11th NASA Goddard Conference on Mass Storage Systems & Technologies, San Diego, April, 2001. (http://www.npaci.edu/dice/srb/Pappres/Pappres.html)

This article provides a peek under the hood of the San Diego Supercomputer Center system that is providing distributed archival storage to scientific communities. The SDSC 'data grid' federates access to a network of diverse but linked storage systems and involves both the Storage Resource Broker (SRB) and the Metacat System. Such a system hides the infrastructure dependencies that arise in establishing user administration while also providing public access to multiple data collections. The design for the grid scales modularly through the use of grid bricks, where each grid brick is a terabyte array of disk drives that create ready access by storing one collection. The scalability technique involves replication of a service by copying both programs and data. Grid bricks are in use at SDSC, under the SRB data management system, serving in the dual roles of primary access and/or back up storage.

At the NPACI URL two related papers can be found: 'MySRB and SRB - Components of a data grid' (Rajasekar et al, 2002) and 'Storage Resource Broker - Managing distributed data in a grid' (Rajasekar et al, submitted). The grid brick concept grows from the disk farm and cyber brick language laid out in Devlin et al, 2002 (http://www.clustercomputing.org/content/tfcc-4-1-gray.html)

Good Read: The Dry and the Wet

- Steven Jackson (PAL) Science Studies Program

Joseph Goguen, "The Dry and the Wet", in <u>Information Systems Concepts</u>, eds. Eckhard Falkenberg, Colette Rolland and El-Sayed Nasr-El-Dein, eds. Elsevier North-Holland, 1992, pp 1-17.

This is a somewhat older piece (first published 1992) that nevertheless raises questions dear to the heart of LTER information managers. The paper points out challenges and outlines strategies for addressing long-standing tensions between 'formal, context-insensitive information' and 'informal, situated information' - the dry and the wet - in the design and maintenance of complex systems. Goguen argues that failures to address this divide (and the general failure to account for the social lives of technology 'in the wild') play an important part in the surprisingly high failure rate reported by complex system

developers. The paper provides a good and accessible overview of the field of Requirements Engineering from one of the leading figures in the field. For links to this and other pieces on similar themes, check out Goguen's website at: http://www.cs.ucsd.edu/users/goguen/.

Good Read: An International Framework to Promote Access to Data

- Don Henshaw (AND)

Peter Arzberger, Peter Schroeder, Anne Beaulieu, Geof Bowker, Kathleen Casey, Leif Laaksonen, David Moorman, Paul Uhlir, Paul Wouters, "An International Framework to Promote Access to Data", in <u>Science</u>. March 19, 2004, pp1777-1778.

The authors recommend open access to scientific data and promote the development of favorable data access policies among scientists, funding agencies, and the national and international research enterprise. While noting there are significant restrictions to open access in terms of security, privacy, intellectual rights, and the lack of appropriate professional reward structures, the potential for increased return from public investment in research is significant. Accomplishing the infrastructure necessary will require continued budgetary support for information management, metadata development, and data repositories and digital libraries within national and international agencies and institutions.