# LTER Databits

Information Management Newsletter of the Long Term Ecological Research Network

## Spring 2011

Welcome to the Spring 2011 issue of Databits! This "bumper" edition of our information management newsletter contains 20 interesting articles, from 16 authors, representing 12 LTER sites.

This is a time of increasing change, with external funding agencies and people both internal and external to our network realizing the value of effective research data management. In this edition, we have evidence that information management in the LTER network is alive and well and making great progress in response to new demands. The volume of articles submitted for this issue demonstrates just how much is going on in the information management arena.

Inside, you will find interesting articles describing network-sponsored workshops; commentaries on a wide range of experiences; descriptions of collaborations between sites; summaries of good external informatics and technology articles; and pointers to useful tools that can support our activities. This is an exciting issue with something for everyone.

In particular, this issue contains several articles from our colleagues in San Diego. Is this because they are clearing out the cupboards? Apparently, it is. We are sad to note that Karen Baker and Mason Kortz are moving on to pastures new. No spoilers (read the article!), but the editors would like to be some of the first, and we are sure not the last, to thank Karen and Mason for their significant contributions to their sites (PAL and CCE), the LTER Network, and especially the IM community. We wish you well in your new endeavors. You will be missed.

Now we know you will want to read on. So, don't delay. Proceed to your nearest purveyor of premium, organically grown, coffee products, with your smartphone, netbook, iPad, or laptop, under your arm. Pause just long enough to order a "moccochoccovanilla, double something" and dive right in!

**----- Co-editors: Philip Tarrant (CAP) and John Chamblee (CWT)**

*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.*

### Featured Articles

### Commentary

### Good Tools And Programs

### Good Reads

### Calendar

### Featured Articles

## Network Identity: 2009 All-Site Milestone and Reflection on Governance

edit

**Karen S. Baker (PAL, CCE) and Nicole Kaplan (SGS)**

This report marks the 2009 milestone of all LTER sites implementing a common access to their LTER website and reviews the process of enactment. Achievement of the milestone means that any LTER site webpage now can be reached by a standardized URL (site_acronym.lternet.edu), e.g., for Palmer LTER, http://pal.lternet.edu/. It is the internet Domain Naming System (DNS) that makes it possible to assign a meaningful LTER Network domain name as an alias for access to each site and aids in representing sites as part of the LTER Network. In addition, the adoption of the alias over time provides an example of the LTER Information Management Committee (IMC) governance at work.

A Databits Newsletter article titled 'Moving Toward Network Identity' (Baker and Brunt, Fall 2001) first presented the notion of creating network identity through agreement on and use of a common network name designation. The implementation of this name-based virtual server was initiated on a volunteer basis. The Web Site Guidelines document (2009 Version 1.1; http://intranet.lternet.edu/im/im_requirements/webdesign_guidelines) delineates three categories of web site elements: 1) design elements that portray the site as part of the LTER Network, 2) organizational schemas that facilitate easy navigation to information and data, and 3) access to current and important data and content. One of the three items mentioned in the first category about contributions to network identity states: "The LTER network domain, lternet.edu, Domain Naming System (DNS) alias (site_acronym.lternet.edu) may be used to reach a site's homepage in addition to the local URL." This Guideline was reviewed and accepted by theNetwork Information System Advisory Committee (NISAC) in 2006 and subsequentlypresented to the LTER Science Executive Board in 2007. In the LTER Site IM ReviewCriteria, reference is made to the 'Guidelines for LTER Web Site Design andContent'.

By the All Scientist Meeting in 2009, the last few sites had implemented the alias prompted by results of an updated survey of implementation and by feedback from site reviews. Though a straight-forward addition of code to a system file for some sites, in other configurations implementation was non-trivial as in the case of a shared institutional web server having existing local conventions embedded in the system file. The implementation by the last few sites frequently required joint efforts of local institution and LTER Network Office system administrators.

The IMC process of implementing the alias over time provides not only the story of an all-site accomplishment but is also a chance to observe a real-life example ofa community-wide, locally-generated example of coordination. The LTER Information Management Committee Governance Working Group has identified such examples as important opportunities to reflect upon the ways of design and/or governance when we are mindful enough to remember, review, and describe our history collectively. This case represents an informal bottom-up approach to designing and enacting a network tool that would identify each site with an LTER web address. Although implementation ran into technical challenges at some sites, it was the compliance of the majority that inspired the last few to make this program a success for the entire Network. This exemplifies a community that voluntarily came into compliance with informal recommendations from the bottom-up as opposed to as a response to a mandate from above. In this case, one might ask whether the formality of a vote would have clarified the use of the alias as an optional guideline or as a network requirement for site reviews. Would an IMC, NISAC, or EB vote clarify agreements? Would their role remain informal? Would any action taken be documented in the minutes of the meeting? Would consideration of bringing it to vote help identify whether this is a topic that would benefit from further discussion? How was the nature of the alias adoption and enactment – unanimous and informal - different from other enacted standards and best practices? Does the alias mechanism represent an expectation that did not cross a tolerance threshold for impacting a majority of local information management workloads? Was it motivated by recognition at the sites of how this would ease day-to-day access to web sites of collaborative partners or was it driven by new expectations from NSF for a more cohesive Network presence online?

The LTER network has reached a milestone of standardized access to each LTER site website. This goal, initiated in 2001 when a coordination mechanism was recognized as both within reach and useful, was reached in 2009. In considering community governance issues and procedures, the milestone continues to serve as a useful vehicle for discussion.

## Making Space for Information Management

edit

**Joan Donovan (PAL, CCE)**

In August of 2010, I undertook an internship with the Ocean Informatics team located at Scripps Institution of Oceanography, where I worked alongside Karen Baker to research the history and development of their Design Studio. The harried halls of Scripps proved a fortuitous space for me to grow as a scholar of Sociology and Science Studies at the University of California San Diego. Science Studies is a program dedicated to bridging the disciplines of sociology, history, communications, and philosophy in an effort to enrich our understanding of the social contexts for knowledge production in science. I had a two-fold objective when studying the Ocean Informatics team. First, I desired to learn more about science-in-action through conversations with information managers, scientists, programmers, and researchers who sought to preserve their data for the long term. Second, I was excited to participate in the work of information management as it developed day to day. In the course of this internship, Karen and I were able to develop and issue a technical report titled: "The Shape of Information Management: Fostering Collaboration across Data, Science, and Technology in a Design Studio."

Using the concepts of cultural theorist Michel de Certeau on space and place, our report describes the history and use of the Ocean Informatics Design Studio located at Scripps through the narratives of those who traverse it. This Design Studio was built with the intended purpose of supporting collaboration among information managers, computer programmers, and scientists/researchers in oceanographic research. However, as demands for digital data grow, so do the purposes for the Design Studio and the responsibilities of the Ocean Informatics team. Through the methods of ethnographic research and semi-structured interviews, we illustrate how the Design Studio works as a piece of infrastructure within this community.

By examining the multiple functions of the Design Studio from the perspectives of a range of users, we found that the space of the studio exerts a paradoxical influence on the collaborative work of information management and scientific practice. While insiders viewed the studio as a dynamic space for collaboration and innovation, outsiders regarded it as a rather conventional place for formal meetings and product demonstrations that nevertheless had an impact on their research practices at sea. This indicates that scientists take the concerns and advice of information managers seriously, especially as it pertains to preparing data and documenting metadata for storage in data repositories.

As well, we found that the location of the Ocean Informatics team on the grounds of Scripps provided ample opportunities for informal meetings in hallways and at lunch tables with scientists and researchers. These impromptu meetings were roundly considered immensely important for OI product development and documenting changes in metadata. Overall, this highlights the need for information managers to stay locally engaged with science so that both disciplines can symbiotically develop and share knowledge.

## LTER Information Management: Continuing Education and Site Change

edit

**Karen Baker and Mason Kortz (PAL, CCE)**

Having garnered an amazing variety of experience and interacted with an equally amazing set of colleagues associated with the LTER data world, Karen Baker and Mason Kortz are leaving their positions in the LTER in order to retool and continue their education. A keynote presentation on environmental policy at an All

Scientists Meetings sparked Mason's interest in returning to law school after six years with LTER. He has decided to attend Boston University (BU) this fall where he has a full scholarship. For Karen, a PhD program in scientific data curation, offered for the first time this fall at the University of Illinois Urbana-Champaign's Graduate School of Library and Information Science presents an opportunity to retire from UCSD/SIO after thirty-five years and to continue her learning while supported by a four-year fellowship.

Information management is carried out jointly at PAL and CCE. Through the work of Karen Baker, Mason Kortz, and James Conners (and Shaun Haber and Lynn Yarmey previously), information management (IM) has, over the last few years, grown into a team effort. Our work with data includes a wide range of site and network activities and provides a unique opportunity for technical, conceptual, and organizational learning and growth. The PAL/CCE team has created a data system that supports many datasets, some of which span more than six decades. We have worked individually and as a team on diverse projects, including the site description dictionary (SiteDB) (1997-2002), the living dictionary demonstration (2005) that later evolved into the current Unit Registry working group, the Web Services Working Group that is currently addressing redesign of the network personnel directory, and the Governance Working Group that led to the development of the IMC Terms of Reference.

Karen Baker was designated 'data manager' for PAL a year after the site began in 1990. PAL scientists concurred that this new LTER 'task' of data management could be accomplished by Karen, who managed and analyzed bio-optical data at the time. In terms of data file size, the bio-optical component -- with its biological, physical, and optical water column measurements -- represented 'the most' data and therefore a member of the bio-optics component seemed qualified to take care of data management. Little did Karen realize when she agreed to take on this task that it would lead to a career change from bio-optical oceanographer to that of an information manager who would become co-director of the Ocean Informatics Initiative and head of a multi-project, multi-agency data repository.

In 2005, Mason Kortz was hired by Ocean Informatics, the PAL/CCE information management group, to begin work on a relational data system that marked a transition from a decade of work with file-oriented data management. Along with James Conners, Mason designed the data and metadata database that eventually grew into the information architecture that is now DataZoo. Later, Mason designed other research databases that bring six decades of biological data into a modern format that can be integrated with current LTER efforts. In 2007 Mason became co-chair of the Unit Registry working group, and in 2009 he organized and chaired the LTER Web Services working group, bringing a site-based design perspective to multi-site and network level efforts.

Having a team approach provides continuity of effort for our information management work. James Conners is stepping up to lead the information management components of PAL and CCE. In order to continue our team approach and collaborative software development, a new programmer will be hired. In addition, efforts initiated by Karen and Jerry Wanetick to provide organizational placement within SIO for information management will continue with restructuring of Ocean Informatics into a recharge facility for information management. Not only does this introduce flexibility in terms of funding arrangements for information management, it provides an institutional identity as well as a resource for researchers requiring data management plans in response to the new NSF requirements for data plans.

The PAL and CCE LTER sites provided a safe harbor for the development of what began as a site dataset commons in 1992 and grew into an information system that now supports four research projects, hundreds of studies and datasets, and a variety of researchers. The LTER requirement for an information manager at each site creates a unique environment, a natural laboratory of sorts, in which the role of information management can grow and develop. The LTER site-based network model provides a unique configuration that both demands and allows for the exploration and definition of the role of information management at a time when scientific research is undergoing transitions in data practice. As illustrated by Karen, Mason, and other LTER IM participants, the network provides a dynamic learning environment for individuals who are interested in careers associated with data and information in the digital realm -- whether it be in relation to scientific research, environmental policy, or digital infrastructure.

## Technical Roles: Am I In IT?

edit

**Shaun Haber (PAL)**

I am sitting in a hotel room in Henderson, NV. This is my home for the next four weeks while I undergo Customer Loyalty Team (CLT) training with my new employer, Zappos.com. I'm taking part in a required training program for all new employees, regardless of position or level, in this customer-service focused company that sells shoes and apparel online. Being temporarily disconnected from my home near San Francisco and dropped into the desert for a month before moving into a new job gives me a space from which to look back over my work roles and experiences.

I recently joined Zappos.com as a Senior Web Developer. Previously, I've held positions at Warner Music Group (WMG; 2007-2011) and the Scripps Institution of Oceanography (SIO) with LTER from 2004-2006. In these jobs, I was also doing web-based work. While these employers vary greatly in their specializations (retail, entertainment, and academic research respectively), but each has much in common for the role of a web developer.

Developers - whether web developers or software engineers - are often confused as being in "IT". For the public, IT is a general term including a whole range of positions relating to computer work, but to an insider this is a naive categorization. Software development and IT are two disjointed worlds. In IT, the objective is the long-term maintenance of systems and status quo. In development, the objective is to build and create. IT focuses on stabilization whereas development focuses on change.

My department at the Warner Music Group was named the "Strategic Technology Group" specifically as a way to differentiate ourselves from the IT department. At Zappos.com, the all-encompassing 'technology' term is dropped completely, and I am a member of the "Development" team.

It is interesting to note that at the intersection between IT-focused systems administrators and software-focused developers is a borderland, the site of an emergent role: development operations, or dev ops. This role acts as a technical liaison between systems administrators and software developers. An effective dev ops lead will make sure software developers produce clean code and applications are deploy-able to production with minimal complications. Likewise, he/she will work with IT to make sure production systems are updated to support innovative products based on new technologies.

Communication is the key across the technological spectrum -- from system administration to dev ops to software development -- for what is highly collaborative team work. In my experience, good communication is frequently enhanced by the creation of documentation. Teams need both a tradition of documenting everything and a shared understanding of the documentation as a reliable reference guide.

After a period of trial and error at WMG, we found that using a plain Drupal site with the book module was most successful as a documentation mechanism. Alternative systems considered included using a wiki, shared documents in Google Docs or Evernote, and README.txt files for software projects. The Drupal book modules organize pages in a hierarchy. New documents were placed in an "Inbox" book and later reorganized appropriately in the "Documentation" book. We relied on Search to find most documents. Categorization seemed to be a necessary formality that we carried out for peace of mind, but in reality we did not actually use it for discovery. Similarly at Zappos, we use a custom MediaWiki installation behind the corporate firewall to house documentation for everything, both technical and non-technical.

Why does this all matter? Regardless whether you are a system administrator, dev ops, or a front-end developer, it's imperative to have a clear understanding of the full technology stack. A well-rounded technology team is more likely to have a solid server infrastructure that supports web and mobile applications optimized for system performance and user experience. Although maintaining good documentation may slow down the development process somewhat in the short-term, it is a small price to pay for enhancing long-term coherence and scalability.

I am looking forward to my new role as a web developer for Zappos.com. Though it is not classified as an IT position in a targeted sense, doing the job well requires an understanding of both IT and Dev Ops. Thus, with a nod to constructive ambiguity, whenever I am asked if I am in IT, I just answer yes!

# Review: The PersonnelDB Design and Development Workshop

edit

**Mason Kortz (PAL, CCE)**

Collaboration in a distributed environment is a cornerstone of the LTER - and of the LTER IMC. This article reviews a collaborative design meeting held in early 2011 to design and develop an updated LTER personnel database.

The LTER Network Office has recently started sponsoring funding for product-oriented working groups – working groups that are focused on the creation of a specific scientific, technical, or organization product. The working groups may be created for the express purpose of creating a product, or they may be formed as a subset of an existing, long-term working group that pursues a general theme across many projects. The LNO funding available to the product-oriented working groups can be used for travel and meeting expenses, providing these groups with an opportunity for face-to-face interaction.

The Web Services Working Group (WSWG) was recently awarded funding for a product-oriented subgroup focused on redesigning the LTER personnel database, and we chose to use our funded meeting time for application design and development. The personnel database represents a continuation of a collaborative model that started with the ProjectDB and Unit Registry projects: applications that are designed and developed by the community, with LNO input, to fulfill a network-level role. The first part of this article describes the meeting process and its outcomes, including both design decisions and applications. The final section reviews the benefits of targeted design and development meetings and considers the role such meetings can play in the LTER network.

## Meeting Structure

The PersonnelDB design and development meeting was held from February 21st to 25th 2011, at the LTER Network Office in Albuquerque, NM. The meeting was attended by six information managers from different sites – Sven Bohm, Gastil Buhl, Corinna Gries, Mason Kortz, Wade Sheldon, and Jonathan Walsh – as well as several members of the Network Office – James Brunt, Mark Servilla, Marshall White, and Yang Xia. The meeting was broken into two segments. The first two days were dedicated to making design decisions; days three through five focused on beginning the development process. All of the participants listed above attended the design portion of the meeting. The development portion was attended by Sven Bohm, Mason Kortz, and Wade Sheldon.

### Days 1-2: Design and Discussion

During the first two days of the meeting, participants were tasked with creating an application specification that could later be used as a reference by the development team. The group began by scoping of the new PersonnelDB database, web service, and interfaces. We discussed potential use cases from the site, network, and public perspectives and determined which of these use cases did or did not fall in the purview of the PersonnelDB. Scoping also included the discussion of which personnel would be included in the database, and who would have access to maintain these records.

As decisions were made and the scope narrowed, the group shifted towards a more technical design perspective. Using the use cases we had previously discussed, the group designed a data model that could represent all of the necessary data and relationships in the personnel database. From this data model we designed two implementations: a relational database schema to store personnel data, and an XML schema to exchange data between the web service and its clients.

Having established design specifications for the data storage and exchange mechanisms, the group worked on more specific technical questions. We created a basic REST syntax for the web service, along with general guidelines for extending the syntax if necessary. We also discussed features for the user interfaces to the personnel service, and mocked up a search interface with a workflow diagram. These were reviewed and compiled with the use cases and data model specifications from the earlier discussions to create a design specification as reference for the development phase.

Throughout the meeting we discussed organizational design as well as technical design. Specifically, we considered the roles and responsibilities of the information managers, network office managers, and network personnel in maintaining the personnel database. The group also discussed the maintenance of the servers on which the development of the PersonnelDB, and future WSWG projects, would take place.

### Days 3-5: Development

The last three days of the workshop focused on developing the PersonnelDB application, using the specifications produced from the first two days. The first task was to set up a development environment in the LNO computational infrastructure; this was done by creating a virtual server and giving the PersonnelDB development team access to manage data and applications on the server. After establishing the server environment, development was done on a MySQL database, PHP entity model and web service implementation, and the XML schema, including XPath and XSLT code for searching and displaying PersonnelDB records.

During the development portion of the meeting, the design specifications continued to evolve as we implemented them. Because of the detail we put into the specifications in the first part of the meeting, most of these changes were relatively small and dealt with technical implementation issues. Although several of meeting participants were not present during the development phase, we were able to continue the design discussion over email and through VTC meetings.

### Ongoing Work

The design and development meeting at LNO was extremely productive, but the PersonnelDB project required work beyond even a very productive week! Part of the wrap-up during the last day was to outline ongoing work and assign tasks to meeting participants. These tasks fell into two general categories: development and documentation. Ongoing development work includes finalizing and testing the PersonnelDB web service and creation of search and management interfaces to the service. Documentation tasks include user manuals, schema and code documentation, and review and analysis of the design process (such as this article). All of this work is supported, in part, by information management buyout time funded by the LNO.

## Meeting Products

The week-long meeting at LNO resulted in several important design decisions, as well as progress on the PersonnelDB application. Here, I describe the most notable decisions made and the reasoning behind them, as well as the application components and documentation that have been, or are being, produced.

### Design Decisions

**Centralized Web Service Enabled Database:** The personnel database redesign was focused on creating an application that was useful in both the network and site contexts. To do this, we decided on a centralized database that would be hosted at LNO, providing a single authoritative source for personnel and roles throughout the network. In order to support the distributed environment of the LTER sites, we also decided on web service access to the data. This allows any site to use the contents of the centralized database as part of their local data system without having to maintain a duplicate copy of the database. For those sites that choose to maintain their authoritative personnel information locally, the web service interface also enables bi-directional synchronization, so sites can easily pull data from or push data to the centralized database.

**Personnel/Profile Database Split:** In scoping the PersonnelDB project, the group divided information about LTER network members into two categories: personnel information and profile information. Personnel information is organizational in nature, and includes a person's roles, site affiliations, contact information, and status as active or inactive within the network. Profile information includes areas of research or technical expertise, working group membership, and collaborations with other LTER members. The PersonnelDB application will handle personnel information, while profile information will be handled by a proposed future application. The profile application would refer back to the PersonnelDB for personnel information.

Because of the organic growth of the existing personnel database, personnel and profile information were mixed together. The decision to split them was not one of the expected goals of the WSWG during this meeting, rather, the decision was made after analysis and discussion of our use cases. Personnel information requires an authoritative source for generating NSF reports, managing contacts for network emails, and integration with other applications. Thus the canonical list of persons in roles at sites needs more control, notification of changes, and archiving of history. Profile information is voluntary, and is more dynamic. Information such as research interests and collaborations will evolve as LTER members self-identify, and so a less rigidly controlled application seems appropriate.

**Multiple Roles per Person:** One of the most universally recognized needs in the updated personnel database was support for multiple roles per person, as the number of individuals associated with multiple sites has increased over time. Discussion led us to identify two types of roles: NSF roles (drawn from a standard list used in NSF reports) and local roles (roles that are unique to the LTER network or even a particular site). The former represents a rigid controlled vocabulary; the latter is more flexible, with managers being able to assign roles freely as well as create new roles. Each person in the database must have one or more NSF roles, and optionally may have any number of local roles as well. The division of roles into NSF and local influenced the management system (see below).

**Management Access:** Part of the organizational design work was discussing the permissions and responsibilities in managing personnel data. The group agreed that role and contact information should be shared between sites to avoid repetition of data when a person is involved with more than one site, but this opened the door for many questions about primary and secondary sites and the management access associated with each one. Ultimately we decided on the following three-tier management scheme:

- LNO managers can create personnel and edit identity (name and primary email) information, assign NSF roles, assign contact information, and designate contact information as primary.
- Site managers can create personnel and edit identity information for anyone with a role at their site. They can also assign and edit local roles and contact information, provided the role or contact information is associated with their site. This means that a site user can add information to any person in the database, but cannot remove or edit information created by another site.
- Personnel will be able to edit their own professional information in the proposed profile application.

**'No Delete' Policy:** In the interest of preserving a record of previous LTER personnel, and previous positions held by current LTER personnel, the group decided on a 'no delete' policy for the PersonnelDB database. Instead, whenever a person, role, or contact information element would be deleted, it is flagged as inactive. This decision was mainly supported by the need for other resources, such as bibliographies, to reference personnel who are no longer active in the LTER network.

**Custom XML Exchange:** When the WSWG members initially discussed XML as an exchange mechanism, we proposed using the Party element from the EML specification. As we compared EML to our use cases, we realized that some of the information we needed to encode, including roles, active and inactive data, and multiple name aliases, could not be represented in EML without extending the schema. Furthermore, the structure of the "Party" element would require repetition of name and identifier information with each role or contact information element. After weighing the complexity of extending EML against the time to draft a new schema, we decided to create a custom schema that more closely mirrored the PersonnelDB data model, but used elements borrowed from EML, allowing for easy conversion back to an EML Party element.

### Application Components

**Web Service:** The primary product developed by the PersonnelDB team is a web service providing access to personnel information. The web service application consists of several sub-components: a MySQL database, a PHP entity model that abstracts access to the database, a PHP service layer that accepts and processes HTTP requests, and an authentication/authorization mechanism that using the LNO LDAP server. The PersonnelDB web service supports both read and write interactions, provides the point of contact between user interfaces and the actual personnel data.

**Web Interfaces:** A second product type is the multiple web interfaces built on top of the PersonnelDB web service. The first interface is a simple search and browse interface that allows users to locate personnel records. This interface is written using PHP and handles all processing on the server, allowing it to be used with a broad range of browsers. Second is a management interface, which allows site and LNO managers to create and update personnel records. Finally, a library of JavaScript widgets is being assembled, including interface elements like person selects and auto-completing text inputs. Combined with the XML tools (see below), these widgets will allow developers to quickly integrate the PersonnelDB into their local sites and applications.

**XML Tools:** A library of XML tools is a third product type. This library can be used with the outputs of the PersonnelDB web service is being created to assist developers in creating their own personnel-enabled sites and applications. These XML tools include XPath queries for common sorting and subsetting actions and XLS stylesheets for transforming PersonnelDB outputs into other useful formats, including EML and HTML.

### Documentation

**Manuals:** User manuals are in development for both the search interface and the management interface. Developer documentation is also being produced. The PersonnelDB XML schema is already fully annotated, with documentation available via the LNO subversion server. A full description of the REST web service interface will also be released for developers who wish to integrate the PersonnelDB service with their own applications. Finally, the data model, MySQL implementation, and PHP code base will be documented to facilitate maintenance and upgrades of the application in the future.

**Reviews:** In addition to technical documentation, the WSWG will be producing documentation of the organizational elements of the PersonnelDB project. The first such documentation was a report made to NISAC in March 2011. The second piece of documentation is this report, providing a summary and retrospective of the design and development meeting. An additional report will be made on release of the service and interfaces. Finally, a living document of bug reports and suggestions will be maintained on the LTER IMC web site.

## Benefits of the Co-located Design and Development Meetings

Beginning a project in a distributed environment like the LTER network can be difficult – without a tangible starting point, participants are often at a loss on where to begin. To counter this, many distributed projects begin with one member assuming the role of project leader and creating an initial prototype product. This provides the group with a starting point, but it also reduces the benefit of having multiple designers. Beginning a design project with a co-located design meeting is one approach to addressing this issue. Personal interaction encourages participants to contribute proactively, rather than just responding to the project leader's work. This leads to a greater investment in the project, which carries over beyond the initial meeting and into the distributed work that follows. In this way, a design meeting can be very important for overcoming the inertia with which distributed projects must cope.

In a distributed environment, focusing the attention of all participants on a single task or issue at the same time can be a daunting task. Because of this, it is often unfeasible to have all members contribute to every decision. This defeats one of the primary purposes of group design – incorporating a broad and diverse set of perspectives. In a co-located meeting, each participant can contribute, and do so in an environment that encourages discussion and dialogue. This leads to more informed decision making, and thus to a more robust design for the project.

Another benefit of a co-located design meeting is the speed with which information is exchanged, challenges are recognized and discussed, and decisions are made. In a distributed environment, participants' schedules often do not align, so collecting group feedback or reaching consensus may take days or weeks. The real-time interaction provided by a co-located design meeting, as well as the lack of distractions, helps focus the group on the task at hand. A design group can accomplish in a few days work that could take months in a distributed environment.

Co-located development provides its own benefits. Initially, it may seem that a day of distributed development work is just as useful as a day of co-located development work. However, development work is rarely the rote implementation of an existing design. As development work proceeds, issues that were not recognized during the design phase may be uncovered or technical limitations may require redesign of the new product. In this way, development benefits from co-location just as design does. Additionally, collaborative development allows for an extremely useful exchange of development methodology in the form of quick, informal code reviews – something that is very difficult to coordinate in a distributed environment.

Overall, a focused design and development meeting is an excellent way to begin a project. The work that can be accomplished in a week, or even in a few days, provides momentum to keep a project moving forward in a distributed environment. Furthermore, the ability to have many contributers working together on the design and development of a project leads to informed decisions and high quality products that can serve as the foundation for future, distributed work

## Commentary

# Systems Upgrade through Technology Transfer across LTERs: Who Benefits?

edit

**John F. Chamblee (CWT) and Wade Sheldon (GCE)**

### Introduction

In 2009, the Coweeta LTER site began planning a complete web and information system redesign. In an early preparatory step, John Chamblee (CWT IM) and Ted Gragson (CWT LPI), met with the Wade Sheldon (GCE Lead IM), to discuss potential use of GCE technology for this effort. Both CWT and GCE LTER sites are administered at the University of Georgia, and attempts to forge closer ties between CWT and GCE have been underway since GCE was first established in 2000. The need for a system upgrade presented a fresh opportunity to push the effort forward. After discussion and demos of GCE software, and with the approval of both project leaderships, we agreed to collaborate on adapting several GCE databases and web applications for Coweeta's use, as well as the GCE Data Toolbox software for processing Coweeta field data.  Although work continues, initial products of this collaboration are now implemented on the new CWT website, unveiled in April 2011 (http://coweeta.uga.edu).

The collaboration between Coweeta and Georgia Coastal has, for the most part, been a one-directional transfer of technology. The Coweeta LTER was able to upgrade information architectures that were out of date with a pre-built system that was more suitable for the present and future needs of LTER information management. If we were to measure the question of "Who benefits?" solely by examining the direct receipt of products by one LTER from another, then the answer must be that Coweeta received the lion's share of return on the investment in collaboration. However, if we add to our evaluation measures of overall product adaptability and long-term potential benefits beyond our two sites, then the question of benefit becomes more equal on both sides of the collaborative equation.

When the question of "Who benefits?" is considered with regard to the broader community, three answers stand out.  In addition to providing a model for future collaborations, this project has introduced some novel approaches to technology that could be adopted elsewhere in the network. We have also learned that there are some additional steps we might take to make it easier to adopt GCE technology in the future.

### Direct benefits for Coweeta

GCE provided a series of applications and databases that were adapted for use in Coweeta systems. The products included, in order of adoption, the GCE Data Toolbox for MATLAB (before it was available under GPL), the GCE_Biblio publication catalog database, the GCE_Metabase data catalog, and the GCE_Submission and GCE_Access databases (which track research projects, additional project resources, and data use). In addition, we were provided an ASP code stack to make views from these databases accessible via the Web. We also developed a shared hosting agreement through which Coweeta could use their LAMP (Linux Apache MySQL Php) server environment to access GCE-originated ASPs via a reverse proxy pass to a GCE IIS server. Finally, Chamblee received many dozens of hours of technical support and mentoring in Information Management. We can break down the benefits Coweeta received from this technology transfer in several ways:

1)  Coweeta did not have to "re-invent the wheel" and invest time and money in a series of new systems, but instead were able to capitalize on existing systems and put into use applications and database designs that had been proven through years of production use.
2)  Since GCE was willing to host CWT ASP pages, Coweeta IM was able to maintain its LAMP architecture and only had to invest in MS SQL infrastructure.
3)  Coweeta was presented with an established and high standard against which to measure their systems. By using our respective data models as a basis of comparison, it was possible to conduct a fine-grained analysis of both data availability and data structure and, in so doing, to document the areas in which Coweeta would need to improve its data management practices.
4)  Coweeta was able to take advantage of a long available but underutilized opportunity for collaboration between LTERs on the same campus, which, in this case, provided the means for informal on-the-job-training. Coweeta experienced a great deal of turnover during the last funding cycle, and Chamblee has only been on the job for two years. By working closely with a more experienced Information Manager, Chamblee was able to gain a broader understanding of the issues that affect Information Management across the LTER Network. In addition, since Chamblee is from a domain science background (archaeology and historical geography), the challenge of adapting GCE's complex information management architecture to Coweeta provided a crash course in several new technologies.

The tools developed by GCE and Sheldon consist of databases grounded on strongly typed data models and highly modular and well-document application code-stacks. This is true of the both MS SQL / ASP applications and the GCE MATLAB Toolbox. This approach to design has unexpected benefits for both incoming and experienced IMs.

While there is no "manual" for most of the GCE applications, per se, the embedded documentation, combined with GCE's modularized approach to design, provides what a linguist might call a "creole, " or hybrid language, documentation of English, various programming languages, and entity relationship symbology. With these tools, a domain scientist who is reasonably comfortable with programming languages and database models can trace out the operational and logical connections and teach themselves the applications. This saves training time for the people providing the technology and it provides an occasional added benefit in terms of helping to locate areas of potential improvement within the technology. By attempting to reverse-engineer programs, the neophyte may also expose an occasional inconsistency in application logic simply by stumbling across a thread that they cannot trace to its conclusion (although these "insights" can also, at times, be due to a simple lack of understanding).

### Indirect benefits for Georgia Coastal Ecosystems

Although the direct benefits for Coweeta were substantial, the somewhat uneven appearance of the cost-benefit balance between GCE and Coweeta is deceiving. In exchange for providing existing technologies and services to Coweeta, GCE received valuable feedback that provided the justification for adjusting or even re-thinking of several design decisions related to GCE's database structures. Moreover, the ease with which GCE's interfaces have been adopted by another site located in a very different biome has also vindicated many of the decisions behind GCE's overall design strategy.

GCE undertook two major revisions of the Information Management system that were at least partially motivated by the feedback Sheldon received from Coweeta. After Coweeta adopted the GCE MATLAB toolbox, Coweeta had some difficulty, overcome with Sheldon's help, adapting the functions that import data logger arrays into MATLAB files.  Sheldon ultimately rewrote these scripts and provided a more modularized approach to a system that was originally designed to handle only the workflows and preprogrammed data logger arrays at the Georgia Coastal site. After this experience, Sheldon took it upon himself to look at several other routines and re-evaluate them for similar issues.

In addition, GCE had long been considering a revision of the entity-relationship model for the GCE_Metabase.  The three principal issues to be addressed were keyword management, instrumentation metadata, and the handling of multi-entity or non-tabular data sets. Before the database was revised, keyword fields were distributed across the database, and tables for individual keyword fields were tied to other fields in parent-child relationships. The new model includes a master keyword table consisting of both keywords and scopes, which define the range of situations for which the keywords apply. Instrumentation was previously handled using a strongly typed model designed to describe the specific domain of instruments that GCE employed, but the new model accomodates simple textual descriptions as well. Finally, a one-to-one relationship between tables and data sets was revised to accommodate many-to-many relationships between data sets, entities (e.g. tables) and files. Data set entities can be time-series data stored in multiple files, a single file containing a multi-table relational database (as with a Microsoft Access Application), or bundles of GIS data in vector or raster formats.

Coweeta's contribution to these revisions is most clear in the case of the data set entity revision. Since Coweeta's IM is a GIS-intensive operation, they were able to provide several use cases against which the data set entity model could be tested. We were also able to provide feedback on the revised ASP applications that took advantage of the new model and, as was the case with the MATLAB toolbox, provide a use case for instrumentation documentation. The difference across biomes between the two sites also allows for broader testing of the new keyword system.

### Long-Term benefits for the LTER Network

The LTER network has the potential to benefit from this collaboration in two ways. First, Coweeta and GCE have demonstrated that it is possible for multiple LTER sites to co-develop a multi-platform web application system and host it smoothly and securely, using reverse-proxy logic to transparently leverage software between sites. Moreover, Coweeta has developed a novel strategy for integrating LTER-specific resources (data catalogs, bibliographies, personnel lists, etc.) with other types of more static content using a low cost hybrid approach that combines proprietary web applications with an off-the-shelf content management system (Drupal).

Rather than developing an entirely new framework using HTML, PHP, or ASP, Coweeta adopted Drupal for managing static resources such as site histories, facility description, driving directions, etc. Once the Drupal pages were in place, ASP-based pages were "re-skinned" to match using include files derived from the Drupal site template. These include files can be edited whenever the site undergoes a large structural change (e.g., a main menu revision), propagating the changes to all ASP-based pages on the site. This hybrid approach, which Chamblee is calling DrASPal, leverages the PHP and JavaScript tools that are included in Drupal and has proven both flexible and reliable.

While the primary participants in this collaboration have been GCE and Coweeta, we have also worked with the Moorea Coral Reef and Santa Barbara Coastal LTERs and included them in our discussions. Together, M. Gastil Buhl and Margaret O'Brien have been working to adopt the GCE ER-models for use at their sites, but in a PostGreSQL framework driven by Perl. As they make progress, they will open up the opportunity for other sites to adopt the GCE database models without any investment in IIS or MS SQL-based technologies.

Moving forward, there are some additional changes that would be worthwhile to pursue. At present, the database instances at each site are referrenced with three-letter LTER prefixes (e.g. GCE_Metabase, CWT_Metabase, etc.). If other sites are interested in adopting these databases, investment in a more generic naming structure might be worth pursuing – especially since sites are likely to continue hosting their own local database server instances. The driving force behind such a revision would take into consideration the cost involved in altering the SQL-based views that stand between web users and the actual database tables. These views often include hard-coded variables, such as server URLs and database instance names. Given appropriate resources, the GCE databases could be restructured in a way that such variables could be stored in tables that are populated when the database instance is first established and that could be easily edited in the event a migration is needed.

In addition, since it was necessary for Coweeta to take a full-scale approach to redesigning our web site, they were able to look over not just their site design and architecture, but also our entire site's content. This opportunity was particularly timely because of the large number of revisions that have taken place recently with regard to NSF policies concerning data citations and data management plans, as well as the proliferation of regulations at universities nationwide concerning research involving human and animal subjects. Coweeta developed their documentation by comparatively examining and citing other sites and LNO and NSF documents and policies. This new documentation provides a good test case for the network and a potential example for other sites to adopt.

### Conclusions

For Coweeta, the list of benefits provided here it is relatively short in terms of descriptive text, but that is because the benefits are so overwhelmingly clear and straightforward. In terms of the time and cost saved relative to developing a new system, these benefits cannot be over-emphasized, nor can we over-emphasize the training value these systems hold for a newcomer.  For GCE, the benefits include feedback on their designs and opportunities to pursue upgrades they were already considering in a context involving much broader use cases. In addition, the success of this collaboration validated the effort and strategic thinking behind GCE's database and software designs, which made it possible to port them across sites and work contexts.

For the LTER Network, the long-term benefit of this collaboration is a model for cross-site technology transfer and development. The GCE /CWT collaboration was successful because the technologies in question were suitable for the purposes to which they were being put and because the principals strove to achieve mutual benefit whenever possible – accepting that the benefits would sometimes be unequal at least in the short term. Over time, this spirit of collaboration is bound to produce benefits that outstrip those already achieved.

## Putting It Out There – Making the Transition to Open Source Software Development

edit

**Wade Sheldon (GCE)**

I have spent a significant portion of my scientific career developing and customizing computer software, both to process and analyze research data, and to build systems to disseminate these data to others. Throughout this time I did what the majority of scientists do, and kept this code mostly to myself. There were many reasons for my closed development approach, from the practical ("the code isn't sufficiently documented for someone else to use") to the paranoid ("I don't have time to answer questions or help people use it") to the proprietary ("why should I give away my hard work for free"). But looking back, one of the primary drivers for my attitude was a negative experience early in my career when I found myself competing against my own software for salary money, and lost. A former research colleague found it more cost-effective to hire an undergraduate student to run my software (developed for another project and shared) than to include me on the new project as a collaborator. Although that issue was eventually overcome, it had a lasting impact on my attitude regarding giving away source code.

After joining the LTER Network in 2000, though, I started interacting with software developers at LNO, other LTER sites and partner organizations (e.g. NCEAS) who were strong open source advocates. Although I still bristled at the thought of sharing my code, the advantages of collaborative software development and sharing scientific software were becoming clear. In 2002 I made a tentative effort at code sharing by releasing a compiled version of a MATLAB software package I developed for GCE data management (GCE Data Toolbox, https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox). Interest in this software was strong, resulting in over 3000 downloads by Fall 2010, but the lack of source code access quickly became an issue. I frequently received email requests for access to portions of the code base for various projects, and reviewers of our GCE-II renewal proposal simultaneously praised us for developing sophisticated QA/QC software and chastised us for not allowing other scientists to inspect the algorithms. During this time I was also encouraged by a colleague to take an online software development course geared for scientists (Software Carpentry, http://software-carpentry.org/), and began to use more community-developed tools for code management and development, including Subversion (http://subversion.tigris.org/), Python (http://www.python.org/) and Trac (http://trac.edgewall.org/). The final straw was an article in Nature (Barnes, 2010) that made an eloquent argument for publishing all scientific software code, warts and all, to allow others to engage with your research.

So in October 2010 I established a policy (with executive committee approval) of releasing all GCE software code as open source under a GPLv3 license (http://gplv3.fsf.org/). As described elsewhere in this issue (Chamblee and Sheldon, 2011), I also began openly sharing database binaries and scripts, web application code and analytical software with other LTER sites (e.g. CWT, MCR, SBC) as well as contributing source code to cross-site software development projects in LTER (ProjectDB, Personnel Database). While I still don't consider myself a staunch open source advocate, and definitely see a place for proprietary software (e.g. to ensure developer compensation, attribution and user support), my experience with "putting it out there" has been fairly painless so far. Requests for help with software code have not increased, and may actually have decreased as people have more opportunity to investigate the code themselves. Although I still worry about getting proper attribution for my work, and cringe when people delve into my code and report back about problems they've found, I know that the software is better as a result.

Although your milage may vary, if you've been holding back releasing your software because "it's not good enough", I encourage you to read Nick Barnes' article and reconsider. If you need extra encouragement (or a fun read), I also suggest looking at "The CRAPL: An academic-strength open source license" (http://matt.might.net/articles/crapl/). My favorite part of the CRAPL is "III. 3. You agree to hold the Author free from shame, embarrassment or ridicule for any hacks, kludges or leaps of faith found within the Program." So please keep that in mind if you delve into GCE software in the future.

References:

Barnes, N. 2010. Publish your computer code: it is good enough. Nature, 467, 753 (doi:10.1038/467753a, http://www.nature.com/news/2010/101013/full/467753a.html)

Chamblee, J.F. and Sheldon, W. 2011. Systems Upgrade through Technology Transfer across LTERs: Who Benefits? LTER Databits: Information Management Newsletter of the Long Term Ecological Research Network. Spring 2001 Issue. (http://databits.lternet.edu/)

## Information Management, in situ: the value of embedding an IM in a Synthesis Working Group

edit

**M.Gastil-Buhl (MCR)**

Moorea Coral Reef investigators commenced a series of monthly “Data Mining Workshops” this year. When I was asked to attend I did not know what to expect.

> *“Data Mining Workshops provide an invaluable opportunity for MCR LTER researchers to come together and synthesize across many different data products. One goal of all LTER sites is that the site should be more than the sum of its component datasets. It is this opportunity to bring together multiple investigators with different points of view and the collection of datasets they bring that provide added value to LTER site science. [These workshops have] catalyzed some important science breakthroughs.” – Andrew Brooks, MCR Deputy Program Director and investigator*

It has been exciting to participate in these workshops and to witness the moment when one investigator sees a connection between their own data and another investigator’s data. Hearing them ask details about each other’s data products gives me clues about what metadata I could add to more fully describe those in the catalog. Several MCR data packages have been enhanced following suggestions made during these workshops.

Our workshop format is participatory, very show-and-tell. We look at each other’s data, all sitting around the same table. Sometimes they jump out of their seats to write on the whiteboard.

> *“Originally we formed these workshops to get more information from our time series, to explore ways to more fully integrate components of the time series program. We saw these workshops as a way to help us address some fundamental science questions as well as bringing us together as a more integrated group. Different from a big science meeting where there are presentations and questions, these are working meetings. Disparate interests can come together in ways that we might not have seen otherwise.” – Alice Alldredge, workshop convener and MCR investigator*

Individual investigators first mined their own data, highlighting trends which might be explained by data from another discipline. From that impetus, collaborations have formed between investigators who had not previously fully taken advantage of the opportunity to work together.

Having an IM embedded in a science workshop helps both the IM and the scientists. I learn how to better serve their needs. I observe which data discovery tools are working well for them and specific ways to improve those tools. The best training for better informatics is to witness the dataset synthesis process. Local workshops are a microcosm of network-wide synthesis projects. Network data synthesis working groups can benefit from inclusion of an IM on their team. The working group gains a participant knowledgeable about data catalogs, formats, and the NIS; the IM gains the broader perspective of network synthesis.

## Notes on Design

edit

**James Conners (PAL, CCE)**

In the last issue of Databits I wrote about the motivations for, and the details of, our current approach to data access within the data system architecture at our site. As described, this approach relies on a web service that accepts the registration of data queries against multiple storage backends. Queries are registered by clients using a standardized query syntax. Clients can then access status responses for the query being processed and data and metadata results are available for access upon successful completion. Currently, we have two applications interacting with this service, DataZoo and an application in use by an education outreach partner. More recently, we’ve had the opportunity to work with a data management group colocated at UCSD/SIO to coordinate the sharing of data through the use of this data access service. From this experience we have obtained valuable feedback for the improvement of our designs.

Realizing from the start that sharing access to our primary web service would risk unwanted increases in server loads and potential negative effects on general performance, we decided to implement an additional service tier that would provide access to a limited set of pre-configured data queries which would in turn be passed through to the underlying service. As developers from the partnering group began testing the API, however, it soon became apparent that additional steps for the mitigation of increased load would be necessary. A mechanism for re-using—within a specified time period—the results for identical queries was implemented. Additional feedback led to other improvements such as improved response headers and updates to general documentation.

In summary, this experience provided us the opportunity to test and improve weaknesses within our design, in addition to assessing its effectiveness. Feedback from collaborators outside our immediate group uncovered issues that wouldn’t have been identified by the limited architectural scope and set of use cases provided by a single data system or group of developers. From my experience, when circumstances allow for it, this type of interaction and feedback frequently provides valuable insight and perspective. Within the LTER network, efforts such as the Tiger Team concept provide a good example of a collaborative arrangement that creates a structure within the development timeline for supporting this type of interaction.

## Information Management, Data Repositories, and Data Curation

edit

**Karen Baker (PAL, CCE)**

Over the last year, opportunities to speak about my experiences in information management have prompted me to consider information management in relation to data curation and data repository efforts.

I attended a Digital Curation Conference in Washington in 2007 and presented an overview of information management within the LTER. I would venture to say one reason it won best paper is because of the uniqueness of the LTER approach and set of experiences with 'data curation'. After spending a decade focusing on the concept of digital libraries, the library community has turned to focus on becoming professional digital data curators. After these meetings, I find the term 'data curation' has entered my vocabulary big time. I began to mull over why LTER information managers do not describe themselves as data curators.

When asked to speak at the Science and Engineering Academic Library (SEAL) California Academic and Research Libraries (CARL) at their Pomona meeting in October 2010, I partnered with a librarian who had been learning from the reorientation of the national library community over the last few years about data curation. My talk, titled 'Information Management in the Wild: A Tale of Local Infrastructuring' introduced the idea of data types as well as of data repository types. Building upon the National Science Board definition in 2005 of three repository types (research, resource, and archive), Lynn Yarmey and I had suggested the concept of a 'web of repositories' at the DCC in Scotland in 2008. After the DCC and SEAL meetings, I find the term 'data repository' occupying a new place in my vocabulary. I began to wonder why information managers who have well-established data collections did not describe themselves as managers of data repositories.

Last week I was invited to speak to a graduate class at the University of California at Los Angeles (UCLA) in the department of Education and Library Science. Having created a two quarter graduate sequence titled "Data, Data Practices, and Data Curation", Christine Borgman begins her materials with a cautionary note from Jeff Rothenberg: "Digital information lasts forever or five years, whichever comes first". The first quarter of the sequence provides a foundation in data practices and services across the disciplines, and the second quarter builds on this foundation to provide practical experience in data curation. My presentation was part of the 'practical experience' section of the class. And it started me thinking about what type of interfaces and infrastructure will be bringing together the worlds of information management, data repositories, and library data curation.

If the local or site-based information management perspective is to be taken into account in the broader arena of scientific infrastructure building, based upon our experience, we site-level information managers are in a good position to speak about data curation and data repositories. We are also in a strong position to discuss design work and collaborative efforts together with network activities and data work processes. As echoed in another article in this issue by Eda Melendez-Colom, the voice of information management is often missing in the stories of data curation and of informatics. Perhaps all those associated with the LTER community can find time and opportunities to speak about the LTER site-based network model configuration in terms of information management. New efforts in larger scientific arenas dealing with scientific data and infrastructure building will benefit if we are able to convey our experiences with data work in practice together with the implications of these experiences.

## Telling the Story of the LTER Information Management: Seizing Opportunities

edit

**Eda Melendez-Colom (LUQ)**

When I was asked by a professor at the University of Puerto Rico to teach a class of her IGERT (Integrative Graduate Education and Research Traineeships) students, I instantly recognized both the opportunity and the responsibility for representing the information managers of the LTER Program. Besides, after years of learning, accumulating experiences and developing knowledge, one feels the need to tell the story the way we have experienced it. It has to do with the knowledge that one acquires by doing, by getting involved, by trying to achieve a goal, and even by doing those day to day boring tasks that most people want to avoid but that we face each day of our lives; an LTER information manager's life. There is a breakthrough that one experiences after 22 years in a job in which every day, week, month and year has felt unique, and where there has been no moment of boredom.

With this as background, there should be no surprise at my dismay when I did not discover a single reference to the role of Information Management (IM) in the story told about the LTER Program in the textbook used for the IGERT EcoInformatics class (Reddy, 2009). There is no indication in this book that such a role even exists.

intertwined memories of events in the history of LTER

I was given full freedom for my presentation to the students so I decided to tell my part of the story. After all, like all the LTER information managers (IMs), I am what they call in history and ethnographic studies, a primary source for many of the LTER events. When I look through the window of my mind, I recall memories of the different stages that LUQ LTER IM has gone through, all intertwined with events of the LUQ LTER program and the LTER as a whole. Figure 1 provides a summary view of some of these events.

I decided that it was only fair to the students (and to all IMs as well) that I tell some of the story of LTER myself as I perceive it, including the important role we have played in the development of informatics in the ecological world.

To that purpose, I selected a paper written by social scientists (those who study how we do our science) in collaboration with LTER information manager Karen Baker (Karasti, et. al. 2006). For this paper, the authors interviewed IMs and participated actively in the 2002 LTER IM annual meeting held in Florida. In their work, they specifically mention the role of information managers in the development of information infrastructure within the LTER program.

In contrast, in the book by Reddy, the LTER Program is rightfully presented as one of the key players in the history of ecoInformatics, but not a single reference is made to the role of information management in it. In the textbook scientists and students are mentioned as the sole managers of their data and although sharing information is mentioned as one of the characteristic of the program, information managers are not mentioned once.

In my opinion, this story is not complete if the role of information management is not included in the narration of the LTER data efforts.

My presentation (Melendez, 2011) contained tables from the book and the paper, comparing first the concepts of the fields of Informatics and Information Management, then the goals, main issues, and solutions to those issues. It is important to point out that there is a distinction between the two fields as made evident by these two sources. While informatics is a science that studies the "design, application, use and impact of information technology", information management is the collection, distribution, organization and control of the information. The former is a science while the later can be considered a discipline.

Table 1 summarizes the various issues related to the differences in the stories presented by the two sources that I exposed to the students. The last part of the presentation was used to familiarize the student with the resources for ecoinformatics that the LTER Program has collaborated in developing, especially the Ecological Metadata Language. In that context, I presented an overview of data handling (Melendez, 2009) and the new tools developed for the production of EML packages, including the metadata online entry forms for the new LUQ LTER IM System. The students were presented with the correct protocol to follow when planning for a scientific experiment in terms of data collection and information management (Melendez, 2009).

**Table 1. Summary of concepts/issues/goals presented by Karasti, et. al. 2006 and Reddy, 2009**

| Subject | Reddy, 2009 | Karasti, et. al. 2006 |
|---|---|---|

| | | |
|---|---|---|
| Goals of Informatics/Information Management | Management and analysis of existing information, large scale ecological research, facilitate environmental research and management, provide common language to computers and humans | Data sharing, interdisciplinary collaboration, large scale-distributed research, archive and preserve, parallel to providing access and reusing data |
| “Impediment”/”Challenge” in obtaining goals | Dispersed and heterogenic information make a need for synthetic analysis tools | Constant intertwined loops of data recovery, entry and archiving current information and planning for the future |
| Solution for the Impediments/Challenges | Create and apply computer technology, developing computer databases and algorithms, automate the access and integration of information | The LTER paradigm: site-network collaboration which places IM as a needed tool for integrating information |

At the end of the class, students were particularly interested in learning more about existing data gathering standards. After conveying the LUQ LTER’s unwritten policy of allowing each investigator choose their own methodology, I pointed out to them of the growing awareness from the scientific community of the need to adopt such standards to ease and/or make possible data synthesis.

I am happy to report the success of this event, based on the involvement of the class and the conversation we engaged with at the end of my talk. Furthermore, my experiences in developing the LUQ LTER Information Management System (IMS), my long time involvement with the LTER Program and the opportunity I have had in working with other scientific communities in the development of their IMS, made it possible for me to give the students a fair account of the important role of information management in the part of the history of LTER and EcoInformatics that I witnessed.

**Acknowledgement**s
I want to thank K. Baker for her willingness to listen to my story and her feedback before and after I presented this class to the students and Professor Mei Yu for recognizing the importance of Information Management and inviting me to address her PhD students in this subject.

**References**

National Science Foundation. August 13, 2003. Fact Sheet: A Brief History of NSF and the Internet. http://www.nsf.gov/news/news_summ.jsp?cntn_id=103050

Reddy, R.A. 2009. Eco-informatics: Tools and techniques. New Delhi, SBS Publishers, 2009. 978-81-89741-99-0 Rs.995.00. (301.31Red/Eco) 078279

Karasti, H, K.S. Baker and E. Halkola. 2006. Enriching the Notion of Data Curation in E-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network Computer Supported Cooperative Work (2006) 15:321–358; DOI 10.1007/s10606-006-9023-2

Melendez-Colom, E.C. 2009. Luquillo LTER Information Management: 2009 Mid–Term Review El Verde Field Station. Rio Grande, Puerto Rico. http://gorilla.ites.upr.edu/files/downloads/luquillo-lter-information-ma...

Melendez-Colom, E.C. 2011. A different story The USA LTER IM case: a presentation given to the University of Puerto Rico INGERT students http://gorilla.ites.upr.edu/files/downloads presentationtoigert04262011.pptx

## Reducing data complications while capturing data complexity

edit

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

The tasks of organizing data for publication are often lumped together as "data cleanup", which implies the process is one of labeling some columns and dealing with some bad values.  In fact though, the process of outlining, clarifying and describing a single data product can require weeks of iteration and communication. Of particular interest to the LTER are time-series data, or data that could be organized into a time series. Often these datasets are comprised of multiple tables, or contain fields that were merged and/or split over time. The longer the time-series, the more likely it is that complications or subtle issues will have been introduced.

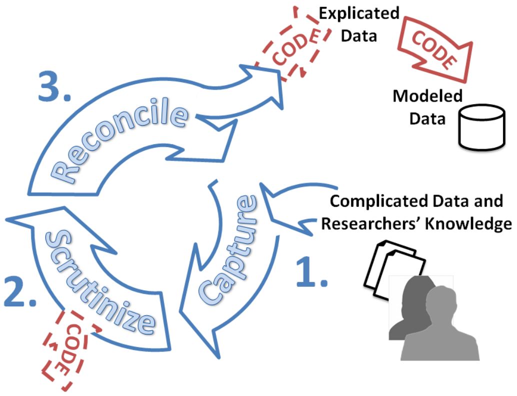
First, a few definitions: from thesaurus.com, something that is **complex** is intricate and subtle, but it also is well organized and logically constructed. But a thing that is **complicated**, in addition to being fundamentally intricate, is irregular, perverse and asymmetrical. “Complex” is more formal and technical (e.g., a mathematics problem) while something like personal life can be “complicated”.  We looked long and hard for a term to use for the process of un-complicating something. The closest we came was “explicate”. Other terms come to mind (e.g., untangle, explain), but these imply a simpler process. **Explication** implies that the subject or thing is more complicated or detailed (also from thesaurus.com).

Ecological data tends to be both complex and complicated.  Complications arise from unavoidable realities that cause missing data or inconsistencies. Instruments break. Cyclones wash away plots. Maintaining a regular structure from irregular input requires coding those complications as missing or flagged values. Quality controls identify complications in the data.  In an ideal hypothetical experiment, there are no complications.  Ecology is complex so naturally the data from ecological observations are complex.  Organisms are classified in hierarchical taxonomy.  Sampling is structured into points along transects within sub-sites within super-sites. Instrument calibrations at a single time point apply to continuously collected data. Capturing that complexity in a model allows manipulations of the data, and in our ideal hypothetical experiment, we still have all the complexity.

Every data cleanup project is unique but they all have some basic activities in common, which we group into two types: data-explication and code-creation.  "Data-explication" is iterative, labor-intensive and mostly manual. It includes extraction and organization of details, and usually solutions are unique. Data-explication will always be required. Code-creation on the other hand is not always necessary; e.g., if an adequate database exists, data might be entered manually. But code can help immensely, and if it’s well planned even ad hoc code can be reused or adapted. Different skills are required for the two types of activities: a good explicator will understand the data issues but may not be able to write adequate code (although s/he must be able to organize details with code in mind). It's unlikely that a professional programmer will have scientific training, and so will not be able to ask the right questions for organization.

We think of the process as having three major steps in a cycle, usually with several iterations (Fig. 1).

Figure 1. The data cleanup process. Cyclical explication steps are in blue. Code-creation is in red, and where optional, the lines are dotted.



**Step 1. Capture the researcher's knowledge.** Usually, this starts with a group of data tables or other files. The researcher and/or field technicians supply information about the data such as the names of the measurements, the sampling methods, and the project's goal and design. Some basic info may be captured with a form, but this is usually followed by discussion. Usually, the older the data are, the more effort this step requires. The asynchronous nature of communication between information managers and researchers (e.g., by email) also adds complications to this step.

**2. Scrutinize the actual data.** Someone must examine the data closely, plotting or tabulating where necessary. S/he looks for inconsistencies in names, categories, data typing, units and precision, and also for shifts in the apparent scope, methodology or ranges of values. One form of scrutiny is to model the data as a relational database schema, and insure that data are clean enough to be uploaded into those tables. Some work can be automated as a workflow or at least a script, e.g., bounds checking, or aggregate counts to verify completeness.

**3. Reconcile the researcher's perception of the data with reality.** Every researcher thinks his/her dataset is less complicated that it actually is. To reconcile, the IM/explicator must be able to ask very specific questions to clarify the discrepancies that became evident under scrutiny. The reconciliation step will result in some data being explained, but also will probably start a new iteration of the cycle.

After all aspects of the data have been explicated, we can move on to the easy part: standardize where possible, re-label the columns and deal with those bad values. If, after this is done, the data can be modeled into a RDB and the tables uploaded, then this is likely to be sufficient. Finally, when the data package is congruent and complete enough to support synthesis, we are done. In general, we've found that the entire process is much simpler when the goal is to add more data to a pre-existing type, than it is when preparing a new dataset for publication.

As data publishers, we've all seen users bypass our data catalogs and go straight to the researchers to learn more about a data set. The network also has the experience of EcoTrends, where one person manually scrutinized nearly all the incoming data, in much the same way as described here. The investment in data cleanup and clarification is particularly important where the goal is to integrate or synthesize. If our data are to be centrally available and in a usable state, then we cannot afford to skip steps or iterations in this process, nor should we underestimate the time it takes. However, it is probably impractical for every dataset to get this treatment, and priorities should be set which reflect our goals and offer the best return on investment.

## Challenges and Opportunities Offered by the NIS

edit

**Margaret O’Brien (SBC) and Don Henshaw (AND)**

In the few months since our 2010 meeting, the LTER Information Management Committee (IMC) has been engaged in multiple projects. Our report to the Executive Board included twelve distinct activities, ranging from very introspective (our Terms of Reference) to co-organizing the upcoming Environmental Information Management Conference (EIMC) for the entire environmental management community. But the bulk of our activities have been associated with the LTER Network Information System (NIS).

The core of our involvement with NIS development has been through the Tiger Teams. Four teams have been active so far: Metadata Management, Workflow Management, Data Management, and Entity Management, with more to start later in the year. Sixteen IMC members have participated or signed up for future Tiger Teams, and we have a volunteer to organize for the group. The IMC put forward six proposals for NIS-related Production Workshops, Training, and Information Manager buy-out time. Following is a summary of recent or planned workshops and training.

Controlled vocabulary: A two-part workshop was organized for the vocabulary; the first, held in early 2011 at the Sevilleta, organized the vocabulary's polytaxonomies, developed vocabulary management with a tool called "Tematres", and planned procedures for modification. A second workshop will take place at the Science Council meeting in May 2011, and engage scientists in defining and clarifying the list. The group's chair, John Porter, also has established vocabulary databases for several individual sites and this group is considering an IMC-wide workshop on creating and managing vocabularies.

Network personnel database redesign: Two workshops and Information Manager buy-out time were funded to redesign and code the personnel databases. Sites will be able to use network database content for their local systems, or to synchronize local systems via web services. The effort is organized by Mason Kortz (CCE/PAL), who is funded through the IM buy-out program to continue this work.

EML Congruency Checker: Work on the EML Congruency Checker Phase I is expected to be completed in 2011 as part of the Data Management Suite of PASTA, and will provide feedback on usability of EML datasets submitted to the NIS. A report on current content in the Network catalog is expected as part of the buy-out award to Margaret O’Brien (SBC).

LTERMapS: this group progressed toward promoting a consistent interactive mapping interface for LTER site information which allows users to visualize, search, download and explore site information. The group held a November workshop at the Andrews Forest which focused on describing GIS data with EML-GIS . Theresa Valentine (AND) is the lead contact for the LTERMapS group, and Adam Skibbe (KNZ) chairs the broader-scoped Spatial Data working group.

Workflows and workflow management: The NIS will make use of automated workflows to process data and create synthetic data products. The first workflows and workflow management workshop will follow on progress made by the Workflows Tiger Team, and will outline Best Practices for NIS workflows, design several example scripts, and develop an agenda for training a small group of information managers as trainers.

DEIMS Training: The DEIMS group (Drupal Environmental Information Management System) requested a training session on installation and configuration of this content management system from LNO personnel. That workshop was held in the Spring, and cross-site work continues on code to develop and present datasets and other site metadata.

The success of production workshops represents additional opportunities and challenges to the sites. As the LTER controlled vocabulary matures, how will sites integrate their local keyword sets with this master vocabulary to assure their site data will be discovered through new search tools? As network databases are redesigned and web services are created, how will sites take advantage of these improved products within their local information systems? A future workshop will consider how ClimDB/HydroDB can be migrated from the harvester approach into the more modern service-oriented architecture provided by PASTA and the NIS. This workshop could similarly address EcoTrends and may reveal metadata quality and completeness issues or data integrity issues that will be ever-present in building synthetic data products. The success of incorporating ClimDB and EcoTrends into the new architecture may be the ultimate proof-of-concept for the LTER NIS.

Then, the success of the LTER NIS will ultimately hinge on the preparedness of site data and metadata. NSF supplemental funds last year were dedicated to information management, and IM-Exec would like to ask each site to describe the progress made using these funds in their annual SiteBite. Demonstration of the usefulness of this funding may be the key to enabling similar offerings in the future. Another assessment of site preparedness will come from checking data sets through the EML Congruency Checker tool.  This year's annual meeting will likely feature discussions on the testing of site data and the best means for reporting on the PASTA-readiness of data for further synthetic uses.

This year's annual IMC meeting and Environmental Information Management Conference (EIMC) promises to be as stimulating as ever, and allows the opportunity for LTER to share its work and ideas with the broader community. This review of workshops and training and their implications for LTER sites indicates there is much to share and discuss. One thing is clear, IMC members and working groups are progressively charging ahead to meet these new challenges.

## Good Tools And Programs

# Wordle: Tool for Generating Text Visualization

edit

**Karen Baker (PAL, CCE)**

Wordle is an application available for free on the web. It creates a word picture of a text where the size of a term is weighted by frequency of its use. Give 'Wordle' a try on a text of your own, it's kinda fun (http://www.wordle.net/). The online interface is straightforward although it allows personalization of the word cloud product via layout, font, and color options.

# Managing Controlled Vocabularies with "TemaTres"

edit

**John Porter (VCR)**

The Controlled Vocabulary Working Group has been working on developing a controlled vocabulary for science keywords to be used in LTER datasets. As part of the process, the working group evaluated software for managing controlled vocabularies and higher level structures (polytaxonomys, thesauri and ontologies).  There are a surprising wide variety of software available, much of it designed for large institutional use, and correspondingly quite expensive (in the tens of thousands of dollars per year).  Needless to say, we focused primarily on low-cost or open-source software. Three major candidates were considered. The "Protege" ontology management software was the most comprehensive, supporting the widest array of lexical structures. However, the large number of features also made it harder to learn and were overkill for the relatively simple goal of developing a polytaxonomy or thesaurus.  The "MultiTes" software is commercial, with a single PC license going for roughly $250 (much more expensive enterprise versions are also available). This had the appeal that it was also used by the National Biological Information Infrastructure (NBII) Thesaurus project. The final software package we evaluated was TemaTres, an open-source, web-based thesaurus management package using PHP and MySQL that came from a library school in Argentina.  Ultimately, the relative simplicity and web accessibility of the interface (while maintaining a rich set of underlying capabilities) of the TemaTres software carried the day.

As we have come to use TemTres for a greater variety of tasks, we continue to be impressed by its flexibility and features. These include a simple, but functional, user interface for editing and browsing, sophisticated search capabilities, and the ability to export all or part of the thesaurus in a number of standardized forms.  Although we have  not yet used the latter capability extensively, the availability of good export forms was a key feature, because it allows us to easily import data from TemaTres into other tools, such as Multitres or Protege, if we choose to at a future time.  TemaTres also supports a more rudimentary input capability for taxonomys in tab-indented files (but fails if their are trailing tabs), or terms in the SKOS XML schema.  The main challenges we have encountered are that some of the menus remain in Spanish, even if English is selected as the language to be used, and that the documentation of some of the more advanced features (such as the ability to link terms between two different vocabularies) is very spotty.

Importantly, TemaTres features a rich set of web services, that provide searching and retrieval capabilities for external programs (http://www.r020.com.ar/tematres/wiki/doku.php?id=:en:tematres:web_services_terminologicos). There are several programs that work quite well with TemaTres, specifically "VisualVocabulary" which provides a graphical view of the structure of taxonomys, "TemaTresView" which provides a JAVASCRIPT-based interface for browsing and "Thesaurus Web Publisher" which provides browse and formatting capabilities coupled to configurable searches using other search engines.

With the help of the LTER Network Office, the Controlled Vocabulary Working Group has established a "vocabulary server" for the LTER-wide vocabulary at: http://vocab.lternet.edu/vocab, including the supporting visualization software (e.g., http://vocab.lternet.edu/thesauruswebpublisher, http://vocab.lternet.edu/visualvocabulary/lter/ and http://vocab.lternet.edu/TematresView/view_thesaurus.php) Some LTER sites have also requested access to the TemaTres software for developing additional controlled vocabularies for site-specific purposes. Fortunately, with TemaTres, setting up a new controlled vocabulary is extremely simple, so that site requests can be easily accommodated (e.g., http://vocab.lternet.edu/vocab/luq).  Additionally, although we are only beginning to fully understand how to use them, there are capabilities for using the web services to link terms between different vocabularies, so that changes to LTER-wide and site-specific vocabularies can be more easily coordinated.  The relatively simple structure of the MySQL database underlying TemaTres also makes it relatively easy harvest keyword from multiple sites for the purpose of identifying new candidate terms for the LTER-wide controlled vocabulary.

# Validating Metadata at the VCR/LTER

edit

**Bridget Long (VCR)**

Metadata is an invaluable tool for information managers. It allows for increased organization and sharing of data all while giving the information mangers the peace of mind that when archived, future users will be able to utilize the dataset. Having a standardized structure for metadata (e.g., Ecological Metadata Language (EML)) has also helped to increase the value of metadata power by improving its readability, flexibility, and utility for archival processing and usage with software applications. These characteristics of EML have allowed it to flourish in the LTER network and in other ecological projects.

Checking the validity of metadata is an important part of the metadata preparation process. Without validation, threats are posed to the future usage of the datasets. Without accurate "data about data", users have difficulty knowing what attributes were measured in the dataset or interpreting the data tables documented in the metadata.

I was charged with the detailed checking of the functionality of EML metadata at the Virginia Coast Reserve LTER, specifically, whether the data tables described in the metadata actually corresponded with the actual data tables. I initially felt overwhelmed at the volume of metadata needed to be validated. However, with the help of a website (http://metacat.tfri.gov.tw/modules/) managed by the Taiwan Forestry Research Institute (briefly described at: http://www.tfri.gov.tw/enu/pub_science_in.aspx?pid=895&catid0=43&catid1=140&pg0=&pg1=1), the task became more manageable. This particular website takes an EML document and creates a statistical program from it using R. It has several different options for how to input EML data, whether by URL or by uploading an EML file and the output can be either in the HTML-R graphical interface or a command line. A third option (that was not utilized in this project, but is worthy of mention) is the creation of a research location from an EML document with Google Maps or Google Earth.

Using the LTER-VCR data catalog, I systematically put each EML document into the website where it would generate an attribute table with each variable, a short description of the variable, the type of variable, as well as its units and range. After generating the table, then a form for the actual dataset would appear. Here, a text file or a comma-separated value file would be needed for input. After that step, either an error message indicating a difference in the data and metadata would appear or the data table would appear in R.

On the whole, the VCR-LTER data catalog of 167 datasets was in good condition. There were only nine datasets whose metadata needed to be fixed and their errors were mainly restricted to the type of variable and its range. Two other datasets needed additional information for one or more attributes. Some datasets, that did not have data tables defined (mostly GIS data), were not checked.

Apart from problems in metadata, I was able to identify certain datasets that were not being displayed correctly in the online catalog. These problems were caused by issues such as broken links and some corrupted database keys resulting from inadvertently added trailing spaces. These issues were resolved using a shell program to correct bad network links and correcting keys using MySQL commands. Without this systematic checking process, these datasets would not have been identified and their problems solved.

*Bridget Long is a student assistant working on Information Management at the Virginia Coast Reserve LTER using supplement funds from the 2010 IM Supplement.*

## Good Reads

---

# Communicating Science

edit

**Jonathan Walsh (BES)**

Steward Pickett wrote a useful summary of the recent Baltimore Ecosystem Study quarterly research meeting. The meeting topic was “Communicating Science.” The summary is online at his web log: http://besdirector.blogspot.com/2011/04/insights-from-bes-communicating-science.html

When it comes to communicating science, the site Information Managers play a significant role. We usually have something to do with the distribution and dissemination of the science performed at our site. We all maintain websites which offer data, news, announcements, and information about our projects. Even if we don’t write much of it, we see it all, and we’re responsible for making it available to the community. Increasingly, the website has become the main method to communicate our research.

Steward’s summary is worthwhile reading for Information Managers. The topics, summarized nicely and presented eloquently and concisely, are:

- Communicating Effectively with Reporters
- Communicating for Policy
- Photography for Communicating Science, Video and Ecology
- Writing for the Public
- Writing For Scientists, and
- Making effective slides.

# A Special Issue of Science on Data

edit

**Karen Baker (PAL, CCE)**

A recent issue of Science (11 Feb 2011) focuses on data and the data deluge (http://www.sciencemag.org/site/special/data/) seems to reveal our nascent understanding of data as a complex entity with multiple scales and with similarities as well as differences across disciplines. The eleven articles in this special issue present different perspectives on data from the areas of climate, ecology, social sciences, health, stem cell, genomics, and neuroscience as well as visualization, signal processing, and metaknowledge. Data issues abound with categories named and frequently identified without the benefit of a broader or shared context. This results in a jumble of ill-described black boxes of data.

Adding to the confusion are statements found in summative articles such as

http://chronicle.com/article/Dumped-On-by-Data-Scientists/126324/

that proclaims “Scientists are wasting much of the data they are creating.” Such a leading sentence is misleading in its over-simplification of complex issues. Buried in this statement is blame and perhaps the suggestion that a choice was made not to pursue an alternative, ie to not waste data. Following the line of reasoning presented that includes assigning blame for data difficulties, the author could also have started off with an equally unsuitable statement that “scientists are taking too much data”.

This issue of Science is important in that it provides a readily accessible aggregation of disciplinary articles, a set of overviews of data in diverse fields. And there are additional hints of new understandings and new ways of knowing. The cover of the special issue itself (see first image below) is not composed of hardware images or streams of 0’s and 1’s suggesting binary bits but rather shows a semantically-summative word cloud. That is, Science Magazine, a science culture barometer of sorts, takes a semantic turn in using a word cloud to convey the notion of data issues. Another word cloud generated in 2008 from a working group abstract on Designing Infrastructure at a Computer-Supported Cooperative Work (CSCW) Conference (see second image below) provides an interesting comparative opportunity. In the NSF articles, the most frequent terms identified are data, research, information, new, climate, science, analysis, visualization, researchers and access while from the workshop studying how we do our science, the top terms include cyberinfrastructure/CI, CSCW, designing, group, research, interoperability, practitioners, collaborative, heterogeneous, and long-term. The former set highlights the what of science while the latter provides a window into issues involved in carrying out science today.

Though this special issue of Science represents a small step toward interdisciplinary understanding of data issues, I remain puzzled at the difficulty in conveying the perspective that those close-to-the-data-origin realize from experience – sometimes called the downstream or bottom-up viewpoint. From this vantage point, there is a lack of appropriate conventions and vocabularies for dealing with the current digital-age capacity to generate data. There is no need to assign blame but rather there is a critical need to recognize that these are remarkable times of transition, a time when we are in the midst of developing new conventions and vocabularies in all fields and at all levels of organization.

## Collaborative, Cross-disciplinary Learning

edit

**Karen Baker (PAL, CCE)**

**Review:** ***D. Pennington (2011 online preprint). Collaborative, cross-disciplinary learning and co-emergent innovation in Science teams.*** **International Journal of Earth Science Informatics*. URL:http://www.springerlink.com/content/81156061q1754t00/***

The contemporary realm of eScience has teams of scientists and technologists working together to create community infrastructure in the form of grids and networks. What enables these teams to collaborate effectively? As made clear in the title of this paper, learning is key. Learning is involved in the establishment of a common ground important for generation of innovative interdisciplinary ideas and for development of shared vision.

Before summarizing, I'll point out that the author, Deanna Pennington, is a member of the LTER community since 1998. Deana conducted her PhD dissertation at Andrews LTER and upon graduation took a postdoctoral position at the LTER Network Office working with the San Diego Supercomputer Center on emerging informatics solutions for ecology. This led to a research faculty appointment, from which Deana participated over the next years in geospatial and remote sensing analysis at Sevilleta LTER; the NSF-funded SEEK informatics project; and in leading two NSF-funded Cyber-Infrastructure (CI) Team projects. She started a new position last year as research scientist in the CyberShARE Center at University of Texas El Paso that allows her to follow her interests in scientific collaboration.

Pennington's analysis introduces a key perspective - the 'learning perspective' – as we strive to work in a manner sensitive to and inclusive of the individual and the larger cognitive system, of training and learning, and of disciplinary and multi-disciplinary work. With the goal of generating shared conceptualizations and ways to develop unified conceptual frameworks informed by semantic issues, the importance of developing not only hypothesized infrastructure solutions but also activities that facilitate group learning and interdisciplinary problem formulation is highlighted. Several models for rapid, co-constructed idea generation are presented that are informed by learning theory.

The conceptual work relating to collaborative learning processes is anchored by observations made in her CI Team project where collaborative team activities were explored. Pennington worked with a variety of multi-disciplinary teams while considering the following: "How does one rapidly develop an understanding between those interested in building tools and those interested in using the tools so that their respective efforts can occur simultaneously rather than being lagged in time?" A very useful table is presented that provides us with new vocabulary and sets of categories for expressing elements of collaborative work. The table makes explicit links between models of collaboration, cross-disciplinary learning and models of technology adoption. More concrete examples from the practices observed would have helped situate the discussion. Fortunately, there are more examples and insights in some of Deana's other recent work (Pennington 2010, 2011).

**References:**

Pennington, D., (2011), Bridging the Disciplinary Divide: Co-Creating Research Ideas in eScience Teams. *Computer Supported Cooperative Work Special Issue on Embedding eResearch Applications: Project Management and Usability*. Online preprint; URL: http://dx.doi.org/10.1007/s10606-011-9134-2.

Pennington, D., (2010), The dynamics of material artifacts in collaborative research teams. *Computer Supported Cooperative Work* 19(2):175-199. DOI: 10.1007/s10606-010-9108-9. [online] URL: http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/s10....

## The Dark Side of the Internet

edit

**Philip Tarrant (CAP)**

**Review:** ***'The dark side of the Internet: Attacks, costs and responses' by Won Kim, Ok-Ran Jeong, Chulyun Kim, Jungmin So. 2011,*** **Information Systems*: 36, 675-705.***

Kim et al. introduce this subject with "The Internet and Web technologies have originally been developed assuming an ideal world where all users are honorable." However, as we now know, not all Internet users have honorable intentions and any individual or organization that accesses the Internet comes under bombarbment from many directions and via many methods.

### Selected highlights:

This comprehensive article takes an in depth look at the different types of Internet abuse from the perspective of: 1) the methods used, both technology centric and non-technology centric, and 2) the damage these attacks cause.

Each type of attack is described in general terms, giving a good introduction to the types of issues that may be experienced.

The authors examine the causes of Internet abuse.

There is a more detailed exploration of how these abuses are implemented and some of the preventative measures that can be taken to reduce risk.

Finally, the authors consider the responses both in terms of legislation and law enforcement that are intended to address the problem.

### Summary:

Arizona State University now regularly scans "asu.edu" web site for vulnerabilities and reports issues to the responsible departments. This process exposes weaknesses that have to be addressed promptly. However, these operations create an overhead for both the institution and the department involved; something, it appears, we just have to get used to. Given the ongoing nature of Internet abuse, for the reader that is less *aux fait* with these different methods, this article provides a good overview of the types of attacks currently in use, the damage they can cause and the attempts by organizations and authorities to combat these issues.

## Calendar

# Events 2011 Summer and Fall

edit

### Event: LTER Controlled Vocabulary Working Group

**Location**: Sapelo Island, (GCE), USA
**Dates**: May 17, 2011

### Event: LTER Science Council Meeting at Sapelo Island (GCE)

**Location**: Sapelo Island, (GCE), USA
**Dates**: May 18-19, 2011

### Event: IMC - Information Management Committee Meeting

**Location**: Santa Barbara, California, USA
**Dates**: September 27, 2011
**Web**: http://intranet.lternet.edu/im/news/meetings/2011

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

### Event: EIMC - Environmental Information Management Conference

**Location**: Santa Barbara, California, USA
**Dates**: September 28-29, 2011
**Web**: https://eim.ecoinformatics.org/eim2011

EIM provides a forum for information managers, scientists, and informatics researchers to present and discuss advances in environmental information management, analysis, and modeling.

### Event: American Society for Information Science and Technology

**Location**: New Orleans, Louisiana, USA
**Dates**: October 9-13, 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.

### Event: DCC - International Digital Curation Conference

**Location**: Bristol, UK
**Dates**: December 5-7, 2011
**Web**: http://www.dcc.ac.uk/events/idcc11

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.