Table of Contents

The FCE LTER Website has a NEW look!

Linda Powell, FCE-LTER

Over the past year, the FCE LTER information management system (IMS) team members, Linda Powell and Mike Rugge, have been working on redesigning and enhancing the FCE website (http://fcelter.fiu.edu/) so that it provides both the general public and scientific community with the types of information and tools that are important to each audience. They released the redesigned website to the public in January 2008. Information is now organized under a series of 'Tabs' and sub-headings in order to facilitate navigation and many of the topical choices are represented by graphical icons as a means to draw the audience into the website (Fig. 1). The IMS team design choices were driven by comments were garnered from the FCE Information Management Advisory Committee (comprised of one member from each of the site roles: PI, Collaborator, Student, Education & Outreach representative and technician), local website users, and LTER information management folks. Time and time again, users commented on how they really did not want to spend much time reading a particular web page but wanted to make their information choices quickly and easily.
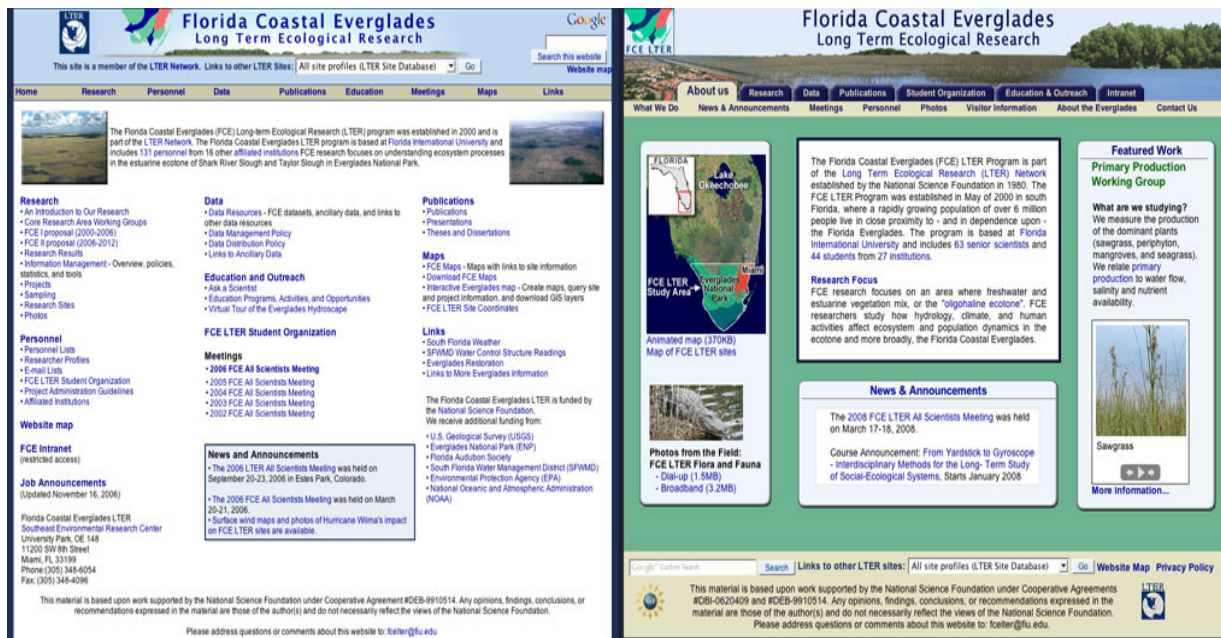
Figure 1. The old FCE LTER home page (left, ca. 2007) and the new FCE LTER home page (right, 2008).

The series of 'Tabs' and sub-headings found at the top of each webpage graphically delivers information at a glance. Because the Florida Everglades has such a rich archive of historical and real-time data collected from numerous state and federal agencies, the FCE LTER IMS team felt it is important to provide their researchers and the general public with links to these essential data. The use of colorful icons and agency logos, particularly in the FCE Data section (Fig. 2), gives the user a means to rapidly navigate through numerous data resources.
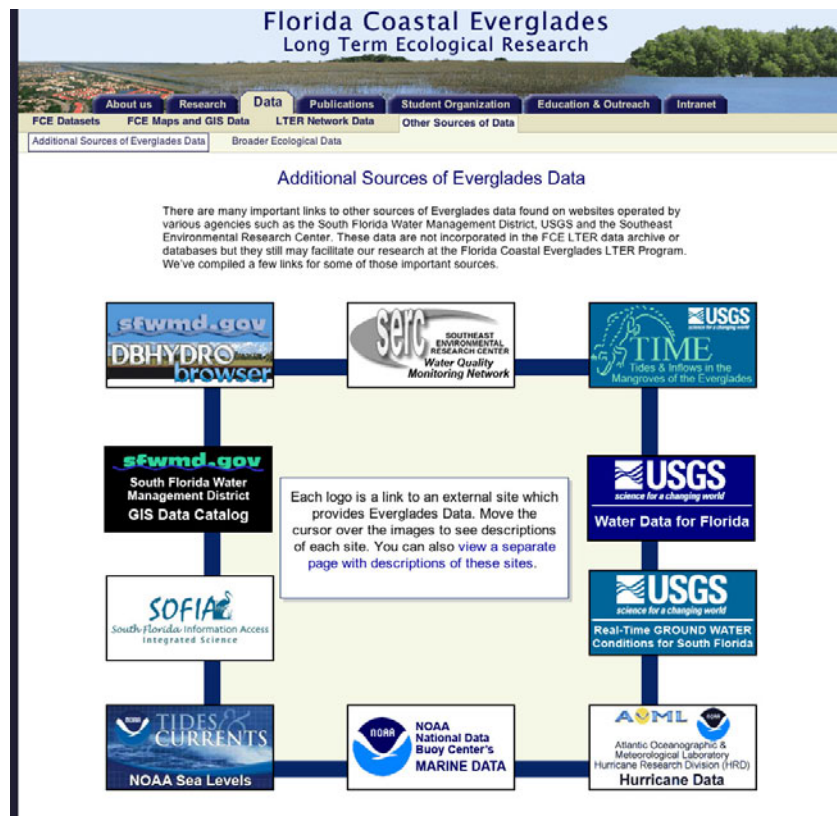
Figure 2. Colorful Agency Logos help web users rapidly navigate data resources.

With the focus of the new FCE website to include more information geared toward the general public, the IMS team partnered with a science writer, Sara LaJeunesse, to translate the FCE research introduction and results into terms that the general audience could understand.

The biggest changes to the website can be found under the 'About Us' tab as the IMS team added features that cater to the general public's interest. The central point of this section is to provide an overview FCE activities and details on the Florida Everglades, such as its history & culture, nature & science, and issues & restoration (http://fcelter.fiu.edu/about_us/everglades/).

The current phase of the FCE LTER program is organized into 5 working groups and 3 cross-cutting themes. The 'Working Group' sub-section (http://fcelter.fiu.edu/research/working_groups/?wg=11&p=FCEII) under 'FCE Research' is a blend of both general and scientific information as there is an 'Introduction' written by a science writer for the public and abstracts, hypotheses and findings written by FCE researchers for the scientific community. Each of the working group 'Introduction' sections also includes how the working group's research affects people in South Florida. Users will also find details and links for all personnel, projects, datasets and publications related to a particular FCE working group.

Figure 3. The FCE website now offers web links to a wealth of Everglades Information.

For the scientific community, the website has many important features, like searchable FCE project information, where research sites, personnel, sampling, datasets and publications are relationally linked to each FCE project (http://fcelter.fiu.edu/research/projects/). Many sections have been simplified and made more user friendly for the researchers. For example, the FCE Data section (http://fcelter.fiu.edu/data/) has a series of graphical icons that represent different data types and sources. Users can select the data type of interest and are taken directly to a data source, such as the LTER Network Data Metacat interface or the FCE core datasets, or given a link to an outside data source like the EcoTrends Project Socioeconomic Catalog. Additionally, FCE core data can now be searched by a theme or by an advance search which includes a spatial search component.

All the relational information found on the FCE LTER website resides in an Oracle 10g database. Information is extracted from numerous sources, like project information forms, researcher EML metadata and publication submissions, and is entered into Oracle Tables. Embperl (http://perl.apache.org/embperl/) and occasionally PHP are used to query the Oracle database and display information on web pages. Changes were made to the FCE website's templating system as DreamWeaver templates were used in the past and the new website uses Embperl's server-side templating system (Embperl::Object). The change to server-side template has made updating web pages easier, since FCE IMS team can make changes to the website without depending on DreamWeaver.

　　　　　Some exciting future website features will include providing the general public sections 'en español' and adding a comprehensive section on Everglades modeling that will not only feature model results, but also perhaps provide a user interface where researchers can run model scenarios through the FCE website.

Developing a Searchable Document and Imagery Archive for the GCE-LTER Web Site

Wade Sheldon, GCE-LTER

In 2007 the Georgia Coastal Ecosystems LTER program began its second cycle of NSF funding, and as part of the transition from GCE-I to GCE-II we conducted a top-to-bottom review of our integrated information system. One major conclusion of this review was that we needed to do a better job of managing the various electronic resource files that are acquired during the course of GCE research and project management activities, including documents (e.g. publication reprints, reports, protocols), imagery (e.g. rendered maps, photos, logos) and other types of static files. During GCE-I, many of these resources were informally organized in server directories using a file system management approach, with online access provided via URL on various public and private GCE web pages. The only effective way to search for some categories of files was using Google Site Search, and many people ignored the web site entirely and contacted GCE IM staff directly for assistance locating specific files.

We also noted in our review that several types of files are already being managed effectively, with file information and network paths stored in relational databases. For example, links to both publicly-accessible and private reprints and presentations are stored in the GCE bibliographic database (http://gce-lter.marsci.uga.edu/public/app/biblio_query.asp). In addition, links to organism photos and other relevant files are stored in the GCE taxonomic database (http://gce-lter.marsci.uga.edu/public/app/all_species_lists.asp). Both of these databases are also integrated with the centralized GCE personnel and metadata databases to support cross-referencing and dynamic linking between personnel records, data sets, publications, and species information. Consequently, we decided to leverage and extend our existing centralized databases and web framework rather than explore the use of other stand-alone file archival systems to provide a more integrated solution.

We began by developing a database schema to store information about files not already managed in databases (http://gce-lter.marsci.uga.edu/public/app/resource_details.asp?id=218). Primary tables are 'Resources', which contains top-level information about resources, such as type, title, and abstract, and 'ResourceFiles', which contains physical file details. These tables are linked via foreign key relationships to lookup tables for type, category and theme, as well as web directory information for URL generation. Resources are also linked to search key words and an authors table, which is actually a junction table to the GCE personnel database. Attributes were also included to support record management operations on dynamic web pages (e.g. 'DisplayOnWeb' bit field in 'Resources', to permit entries to be taken offline), as well as file-type-based and custom thumbnail images for each resource ('IconURL' in FileTypes and 'Thumbnail' in ResourceFiles, resp.). The database schema was implemented using SQL Server 2000, and SQL views were developed for searching and displaying database contents, as well as to dynamically integrate information for general file resources with information for reprints and publications in the GCE bibliography and images and other files in the GCE taxonomic database to support transparent cross-database queries and record display.

After populating the database tables with information about resource files stored on the GCE file server (using a combination of data mining and manual entry approaches), we developed dynamic web applications for querying this database and displaying file details (http://gce-lter.marsci.uga.edu/public/app/resource_search.asp). The main search page contains a top panel with drop-down menus for selecting file type, category and theme, and text boxes for specifying search text in the title, abstract and key words and author last name (fig. 1). Record display option fields are also provided for specifying sorting, abstract display and records per page. Below the search panel is a dynamically-generated "browse" interface, displaying the contents of the database listed hierarchically by type, category and theme. Clicking on any level in the hierarchy executes a search for files matching the corresponding terms automatically.
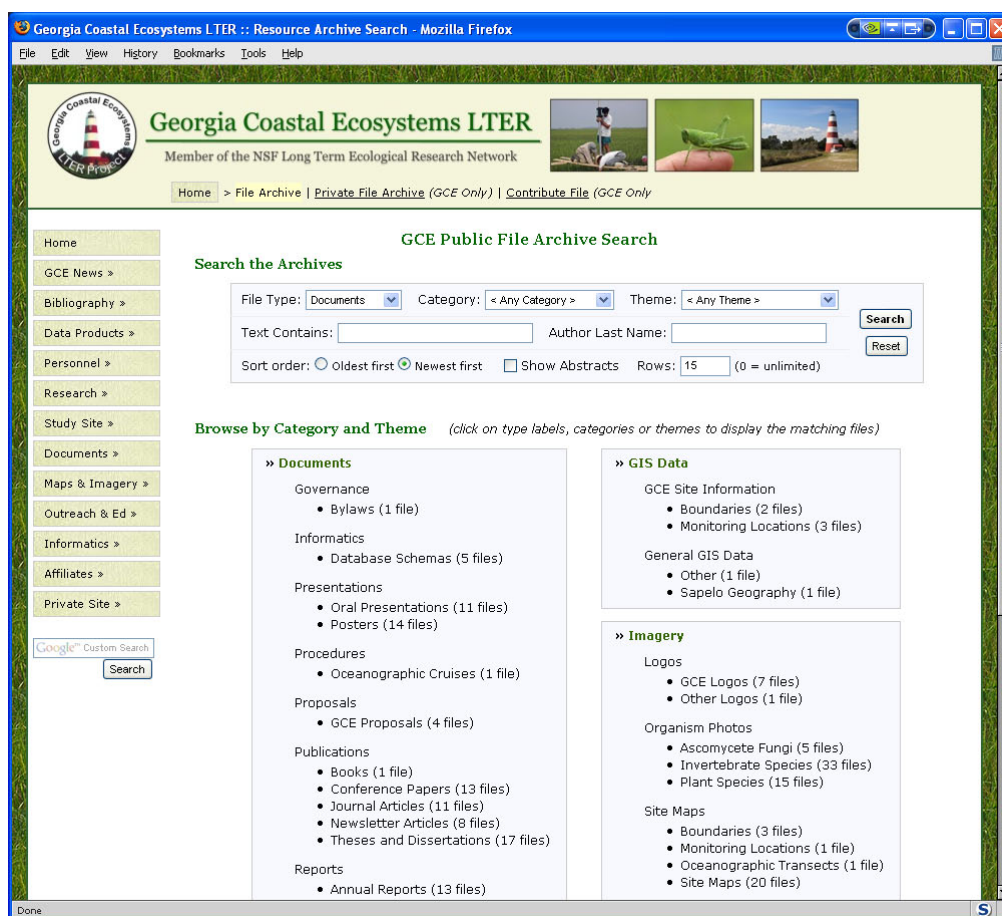
Figure 1. GCE document and imagery archive search and browse interface

Search results are displayed in a summary table, with results grouped by category and theme (fig. 2). Complete titles are displayed along with contributor name, year, and abstract. Display of abstracts can be controlled using options in the search panel and also toggled by pressing the "Abstracts" or "Hide Abstracts" button at the top of the page.

Clicking on the file icon or thumbnail image downloads the file directly, as indicated by the link help text (displayed by hovering on the image). Clicking on the title opens a document details page that includes all available information on the file, including title, abstract,

Figure 2. GCE document and imagery archive search results page

contributors, file date and version, download link and file size (fig. 3). The corresponding virtual directory hierarchy is also listed in the "Archive" field, with each entry hyperlinked to execute a query for documents matching the respective type, category and theme. Corresponding breadcrumb navigation links are also provided for parity with the rest of the GCE web site. A citation is provided for every resource, which is either drawn from the bibliographic database (i.e. for reprints and presentation files), or generated based on authorship, origination date and title information in the resource database of taxonomic database for non-bibliographic entries.

An identical set of web applications is also available on the private GCE web site for project participants. These applications support searching and browsing for all resources in the public archive as well as many additional files that are not publicly available, such as restricted-access reprints, project governance information, unreleased presentations from project meetings, and confidential personnel information. The distinction between public and private resources is controlled using the "PublicAccess" bit field in the "Resources" table, and reinforced using appropriate web directory access permissions on the server. GCE participants can use web forms on the private site to archive new files and update prior entries, with access control enforced based on project role and login so that only the original contributor, a co-author, or site IM can revise prior archive entries. JPEG thumbnail images are generated automatically on upload for supported image file types using a server-side thumbnail generation component (ASPThumb).

Figure 3. GCE document and imagery archive resource details page

The new GCE document and imagery archive provides many important benefits to web visitors and project participants. First and foremost, **all** public and private file-based GCE resources can now be archived, discovered, and accessed using an integrated web-based interface. Stable URLs and full citations are also provided for all resources to support external hyperlinks and appropriate attribution for GCE content providers. Thumbnail images are also provided for all imagery, including maps, site and organism photos, and logos, to permit visual browsing for items of interest.

This archive also simplifies file management and web content management for IM staff. GCE personnel can now archive files on their own, and the dynamic cross-referencing of bibliography reprints and publications, as well as species list photographs, alleviates the burden of maintaining information about these resources in multiple databases and web pages. File versioning is also handled automatically, with date/time stamps appended to file names on upload to prevent over-writing of prior versions. URLs for general or specific archive queries or resources can also be included in web navigation menus and page links, providing direct access to frequently-requested content. Additionally, recent additions to the archive are automatically listed on the dynamic GCE news page (http://gce-lter.marsci.uga.edu/public/app/news.asp), ensuring that visitors are aware of new uploads and updates of interest.

Improving the Basic Keyword Search for Datasets by Employing Text Mining Techniques and Indexing

Hung V. Nguyen[1], Corinna Gries[2], and Hasan Davulcu[1].

[1] Department of Computer Science and Engineering, Ira A. Fulton School of Engineering, PO Box 878809, Arizona State University, Tempe AZ 85287-880

[2] Global Institute of Sustainability PO Box 873211, Arizona State University, Tempe AZ 85287-3211

Introduction

Most datasets at Central Arizona Phoenix LTER are now well documented in the Ecological Metadata Language (EML) and are available on the web. By strictly applying an internal, limited controlled vocabulary these datasets are organized into categories by which they can be browsed on the web. In addition they can be discovered by searching for the authors and the specific project which produced these data. Initially a keyword search and a full text search engine were made available to the user. However, usability testing quickly showed that search success was very limited. Generally, this problem is based on the fact that the particular keyword a user might be searching on is not occurring in any of metadata files although some datasets are relevant to the general subject the user is interested in.

To bridge the gap between search terms and terms used in describing the dataset in the metadata file a search engine was developed which combines free text search with suggested similar keywords, i.e. a more guided search providing a higher success rate. Text mining techniques were applied to EML files to extract terms and compute strongly correlated sets of terms. The terms are then indexed for faster search performance and the user is provided with immediate feedback on the number of hits.

The human evaluation shows that our approach yields high accuracy in terms of relevance for both related keywords and discovered datasets. In addition, we developed a system in which an administrator can add new EML files to the search engine. Each new file will be automatically mined for relevant terms and the terms integrated with the similarity matrix. The administrator also has the option of editing the sets of terms for their human perceived usefulness in searching for ecological data.

Data Extraction and Dimensionality Reduction

In data mining, the data preprocessing and data cleaning steps are very important. Preprocessing data affects the quality and the efficiency of the data mining algorithms. Initially we obtained about 25000 terms from abstracts, titles, and keyword lists in EML files stored in the LNO Metacat resulting in a very large term-document matrix. The number of rows of the matrix M corresponds to the number of documents in the collection and the number of columns corresponds to the number of distinct terms with stop words removed. Each matrix cell holds the term-frequency (TF, frequency of term in document).

In a second step the term-document matrix is used to define a term – term correlation matrix containing normalized correlation values between terms in the collection. Two terms are correlated if they have high co-occurrence in the documents. The size of the matrix C is m x m, where m is the number of terms in the collection. As a consequence, the dimensionality of the matrix C is still very large and it is necessary to reduce the dimension. Many rare terms do not contribute significantly to the correlation matrix. To reduce this 'noise' strongly correlated term sets were constructed by computing co-occurrence clustering. Co-occurrence clustering helps reduce the dimensionality of terms and also increases the correlation degrees within the clusters of terms that will be used as term sets for association rule mining. More specifically, we employ the Hierarchical Agglomerative Clustering (HAC) technique to do the co-occurrence clustering of terms. At the beginning, each single term forms a cluster. Subsequently, we try to merge two *closest* clusters. The clustering process stops when the cardinality of merged cluster exceeds a threshold t. There are several ways to measure the distance between clusters. In this work we followed the UPGMA (Unweighted Pair Group Method with Arithmetic mean) scheme. UPGMA is a simple agglomerative or bottom-up data clustering. The algorithm examines the structure present in a pair-wise distance matrix to then construct a rooted tree (dendrogram).

This process eliminates rare terms and therefore reduces dimensionality, after which the document term matrix is re-created for each of the term clusters. This matrix then is used for term weighting in the search engine.

Metadata Document/Datasets Retrieval

Each obtained cluster contains highly correlated terms and the cardinality of each cluster is much smaller compared to that of the whole collection of terms. For each cluster a hash table is constructed which in turn is used to build a vector of elements in that table and compute the cosine similarity between this vector and EML documents in the collection.

We have implemented a web-based search engine prototype to test the above outlined technical approach. The software consists of several components:

- The indexing component for dimensionality reduction and correlation computing
    - Construct Term-Document (term-doc) matrix M:
        - $M_{i,j}$ = frequency of term i in document j
    - Compute normalized term correlation matrix S =M x M'
- The search and retrieval component
    - Process the user's queries
    - Return the related keyword list to each query term (using indexing engine)
    - Return the relevant datasets
- The management component that allows an administrator to add new EML files
    - Load EML document into eXist XML database
    - Extract important terms from new EML files
    - Allow administrator to edit list of terms
    - Update term correlation matrix and index

Future work

Currently we are researching ways to further bridge the gap between concepts a data user may search on and terms used in EML files to describe the data. For this purpose we have included related publications, ecological glossaries and text books, and the preliminary LTER controlled vocabulary into the data mining and indexing process. This increases the number of related terms a user is presented with. However, not every term will lead directly to the discovery of a dataset. Therefore, more research is necessary to automate the process of developing a hierarchy of terms in which more general terms will be linked to more specific terms that are found in the EML files and therefore will lead to the discovery of a dataset.

Acknowledgements

Developing and Using APIs in System Design

James Conners and Mason Kortz, PAL-LTER/CCE-LTER

An Application Programming Interface, or API, is a set of code that abstracts complex operations into a simple interface. The interface provided by an API is not intended for use by end users, but rather by developers, as one way to incorporate existing functionality into an application while minimizing the amount of learning and coding necessary. Well-known APIs include the Google Maps API, which abstracts complex database calls and browser rendering into relatively simple Javascript functions, or the many database APIs that abstract the opening of sockets, sending of queries, and structuring of results into a small number of calls in your favorite language. In early 2007, the Ocean Informatics team began exploring APIs for various needs such as in-browser plotting, database abstraction, and XML transformation. As our in-house use of APIs grew, we began to organize our own code into APIs, and in November of 2007 we finished created a personnel management database schema with an accompanying API definition (and PHP implementation). Subsequently, we have created APIs for terminology lists and media collections, and have begun work on an API for the LTER Unit Repository.

Using APIs as part of the development process provides several benefits. By utilizing existing code, developers save time that would be spent generating and testing thousands of lines code. Some APIs provide interfaces into technologies so complex that it could take years for an individual or small group to recreate them - as is the case with the Google Maps API or the Microsoft DirectX API. In this way, APIs enable the development of richer applications on shorter time scales. Another benefit of using APIs is that the developer only needs to learn the interface into the technology being supported, not the technology itself. Database interface APIs, for example, allow developers to create networked database applications without knowing the ins and outs of sockets, message headers, and responses. The incorporation of high level modularity results in applications that are more easily shared and extended because of their transparent compartmentalization of logic and functionality.

In addition to the benefits of using existing programming interfaces for application development, there are also reasons for using APIs as a local development model - producing local work based on core code bases and storage resources with associated APIs. One possible benefit of this model is the resulting implicit community standards of development practices. API-oriented development begins with a defined set of capabilities required for an interface to be useful. This set of capabilities includes both immediately necessary functions and functions which seem likely to be useful in the future. Because the goal of API development is code reuse over time by developers within or across communities, documentation and clear code are encouraged, being primary factors in the useability of a programming interface. Having a well-defined, documented interface into a resource allows developers throughout the community to utilize them without concern for back-end changes deprecating their code. These factors result in a community of developers producing a well-documented and structured core of shareable works that provide enough stability for implementations built upon them to maintain a degree of flexibility and be developed with agility. These core resources are consequently well suited for sharing within or across development communities.

API-oriented development practices have both positive and negative ramifications. One such trade-off is the loss of rapid initial development periods in favor of more efficient program implementations in the future. More time in the beginning phases of development is required, but the benefit of this is a facilitative code interface that provides a substantial base structure on which to build additional applications, each providing access to a shared technology. An example would be an API developed for a media gallery backend that stores and queries photos, movies and documents. Implementation of an application using the media gallery technology begins with a large part of the development already completed. The application development cycle is limited to providing end-user access to the existing capabilities of the API, cutting development time substantially. Implementations can be across platforms, each with a different look and purpose, while all relying on the same code base. The benefit of this trade-off, then, depends on the anticipation of multiple applications using a common code base.

Code abstraction is another implication of API development, with the granularity of program control strongly influenced or restricted by the design of the programming interface, establishing a lower bound for optimization. APIs may provide interfaces to resources that include modularized storage which may affect performance of data retrieval or enforce weak relational references across databases. Additionally, abstracted code is less fluid than procedural code, as the API developer is committed to encapsulating the functionality of a code base in a set of established functions. Abstracted code is also much easier to read, and an abstract interface into a complex code base has a much shorted learning period than the code base itself. APIs are more easily adopted into new or existing development cycles, whether within or outside of the community that developed the API.

API-based development also entails certain changes in development practices. A longer period of prototyping before coding begins is beneficial to an API, as it allows the developer to consider future uses of the interface beyond immediate needs, and to discuss the requirements of the interface with other developers who may use it. An organized code repository, with agreed-upon standards for quality and completeness, increases the mobility of APIs, working towards the goal of community reuse. As emphasis on code reuse increases, so does emphasis on documentation - because abstracted code is not self-describing, time must be taken to describe it clearly for it to be useful beyond the immediate developer.

The decision to develop an API frequently occurs when a technology or resource exists from which one or more communities would benefit. Providing a programming interface into this resource encourages not only the sharing of code but the usage of a common technology that creates in effect a coordinating mechanism that makes more solid the basis for collaboration. Efforts across communities combine where they would normally conflict. Though not all undertakings' criteria justify the need to develop with such a scope, we have found that many of our local projects have benefited substantially from an API-oriented development model, both functionally and organizationally.

FinLTSER Network: New addition to the global LTER network

Helena Karasti, FinLTSER (ILTER)

The FinLTSER Network (Finnish Long-Term Socio-Ecological Research Network) was established by a decision of the high level Coordination Group for Finnish Environmental Research in December 2006, on the basis of an international evaluation. FinLTSER was formally accepted as a member network in ILTER and LTER-Europe at the ILTER meeting in Beijing in August 2007. The recently established European LTER network (LTER-Europe) that was formed through the merger of existing more regional European networks is now being consolidated in the EU Network of Excellence ALTER-Net (A Long-Term Biodiversity, Ecosystem and Awareness Research Network) for biodiversity and ecosystem research. FinLTSER also participates in the LIFE-WATCH initiative (e-Science and Technology Infrastructure for Biodiversity Data and Observatories) that was selected to continue into the preparation phase among the most promising next generation large-scale Research Infrastructures by the European Strategic Forum for Research Infrastructures (ESFRI).

Table 1. Acronym list with associated names and links.

FinLTSER	Finnish Long-Term Socio-Ecological Research	http://www.environment.fi/syke/lter
LTER-Europe	Long-Term Ecosystem Research and Monitoring in Europe	http://www.lter-europe.ceh.ac.uk
ILTER	International Long-Term Ecological Research	http://www.ilternet.edu/
ALTER-Net	A Long-Term Biodiversity, Ecosystem and Awareness Research Network	http://www.alter-net.info/
LIFE-WATCH	e-Science and Technology Infrastructure for Biodiversity Data and Observatories	http://www.lifewatch.eu
ESFRI	European Strategic Forum for Research Infrastructures	http://cordis.europa.eu/esfri/

FinLTSER consists of nine LT(S)ER sites and research platforms, representing the main ecosystems in Finland, including marine, terrestrial, lake, sub-arctic, urban ecosystems. The Finnish Environment Institute (SYKE) acts as coordination body of the network. The FinLTSER Network is formed by research stations of the universities of Helsinki, Jyväskylä, Oulu, and Turku as well as research sites, instrumentation and long-term monitoring programs of some of the main governmental research institutes, i.e. Finnish Environment Institute, Finnish Meteorological Institute, Finnish Forest Research Institute, Finnish Game and Fisheries Research Institute, and MTT Agrifood Research Finland.

The core research areas of FinLTSER align with the ILTER core areas of global water circulation, biogeochemical processes, and changes in biodiversity. In addition, FinLTSER research areas emphasize: other ecosystem processes and disturbances; ecosystem services, e.g. soil health and soil fertility, forest and agricultural production, water resources; societal and other socio-economic pressures on the functioning of the ecosystems; effects on the local communities of nature conservation and resource exploitation; and local environmental

conflicts. Relevance of the 'social' component (Haberl, Winiwarter et al. 2006) is self-evident in these particular research areas and also in the FinLTSER Network's composition, i.e. four out of nine sites and platforms are LTSER.

The FinLTSER Information Management Group (IMG) was formally approved by the FinLTSER Steering Group in April 2008. The IMG comprises participants responsible for information management at each of the LT(S)ER sites and platforms together with representatives of associated research institutes. Two representatives of the Information Management Group, co-leads Helena Karasti and Pirjo Kuitunen, were accepted as members of the FinLTSER Steering Group. The IMG held its first meeting in University of Oulu in March 2008, and is currently planning the second one to be held in Konnevesi research station (Lake Päijänne LTER) in May 2008. The IMG co-leads have also started networking with European and international LTER networks. Helena Karasti and Pirjo Kuitunen are members of the ALTER-Net Work Package I6 called 'A framework for effective information and knowledge management'. In addition, Helena Karasti is a European member in ILTER Information Management Committee together with David Blankman (LTER Israel) and Cristiana Cocciufa (LTER Italy).

A survey was organized in December 2007 - January 2008 to review the information management capability within the FinLTSER sites and platforms. Finnish LT(S)ER sites house a wealth of national ecological data heritage. As FinLTSER Network has, to a great extent, been built on the existing network of research stations, the sites have up to date managed these legacies according to the needs of local research. Currently coherent plans and specialized personnel for information management are lacking which compromises the integrity and long-term availability of the legacy data. Entering the global LTER program and the e-era of LIFE-WATCH initiative this data management approach is not sufficient. Therefore, a site-based network approach to information management needs to be developed so that it accounts for the particular problematics of both the long-term socio-ecological research and the e-era worldwide research enterprise that it aims to support.

This development can be seen as two major, partially overlapping phases that align with the recently forwarded trajectory for ILTER sites' information management: 1) development of LTSER information management strategy and system aimed towards e-infrastructure (a.k.a. cyberinfrastructure in the US) capability in accordance with international policies, standards and planned development trajectories (ILTER, LTER-Europe, LIFE-WATCH), and 2) development of e-infrastructures to allow multi-site comparative analysis and experimentation, modeling and scenario development, integrative socio-economic research as well as participation in European and global LTER research initiatives. Many of issues FinLTSER IM faces in the context of a global research network can only be meaningfully addressed, successfully developed and eventually achieved through international collaboration, and, in fact, this collaboration began already some years ago.

Some of the US LTER Databits readership may remember me from the year 2002 when I participated in the NSF funded work on BioDiversity and EcoInformatics (BDEI) project called "Designing an Infrastructure for Heterogeneity in Ecosystem Data, Collaborators and Organizations" that was lead by Karen S. Baker and Geoffrey C. Bowker. I attended many meetings, made a number site visits and conducted over 50 interviews with US LTER participants. Furthermore, Karen Baker (PAL, currently also CCE) and I presented our initial findings in a talk "Bringing Everyday Practices and Lived Experiences into the LTER Metadata Discussion" at the 2002 LTER IMC in Orlando, FL. Let me use this

occasion to sincerely thank all the US LTER participants who so generously offered their time, experience and expertise and thus gave me a unique opportunity to learn about the long-term perspective, site-based scientific information management, and the issues associated with large-scale scientific collaboration. The materials collected have been used in analyses that have culminated in several publications about the US LTER information management (Karasti and Baker 2004; Karasti, Baker et al. 2006). The materials are still in use and our collaboration continues (Karasti, Baker et al. 2007; Karasti and Baker 2008).

Upon returning to my home country of Finland to an information systems department at the University of Oulu, I learned that Finland was in the process of becoming an LTER network. Therefore, it was an easy decision to continue with my awaken interest in LTER related issues in several ways. I formed a Masters thesis group in which students used the materials gathered within the US LTER Network. Students' topics ranged from identifying different views on information technology, conducting narrative analysis in order to trace the collaborative development of metadata, addressing the problematics of open data access and use policy in scientific research, to using the notion of information infrastructure in analysing scientific collaboration within the US LTER Network.

My own current research interests focus on elaborating upon the long-term perspective and problematics, particularly in the context of building information systems and infrastructures, combining participatory design and ethnographic approaches which has been my long-term interest. As FinLTSER was being established, I have drawn on my introduction to and study of the US LTER approach to information management in order to work with the network in formation in Finland (see above). We have also started to gather longitudinal data about how information management is developed within the nascent FinLTSER Network. We have had a student project that began to develop a site level framework for LTSER information management through an empirical study of one of the LTSER sites within the Northern LTSER platform. I currently supervise one doctoral and two Masters students who have LTSER related topics and carry out empirical work within the FinLTSER Network. Furthermore, I continue to interface with US colleagues and the US LTER Network. ILTER (KristinVanderbilt) and the IMC (Karen Baker) have provided continuing support in various ways that share US insight and experiences with FinLTSER IMG participants. I am convinced I speak for the entire FinLTSER IMG in saying that we are looking forward to all future collaborations relating to LT(S)ER information management and e-infrastructure (cyberinfrastructure) development.

References

Haberl, H., V. Winiwarter, et al. (2006). From LTER to LTSER: Conceptualizing the Socioeconomic Dimension of Long-term Socioecological Research. Ecology and Society, Vol. 1, Issue No. 2, Art. 13.

Karasti, H. & K. S. Baker (2004) Infrastructuring for the Long-Term: Ecological Information Management, Hawaii International Conference on System Sciences 2004 (HICSS'37). January 5-8 2004, Hawaii, USA, pp. 10.

Karasti, H., K. S. Baker, et al. (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 – An International Journal, Vol. 15, Issue No. 4, pp. 321-358.

Karasti, H., K. S. Baker, et al. (2007) Digital Data Practices and the Long Term Ecological
    Research Program. 3rd International Digital Curation Conference. December 11-13,
    2007, Washington DC, USA. pp. 13.
Karasti H. & K. S. Baker (2008) Digital Data Practices and the Long Term Ecological
    Research Program Growing Global. International Journal of Digital Curation, Vol. 3
    Issue No. 2 (in press).

Contact details:

Helena Karasti (FinLTSER, Northern LTSER Platform)
Department of Information Processing Science, University of Oulu
P.O.Box 3000 , FIN-90014 Oulu University, FINLAND
email: Helena.Karasti@oulu.fi

Using Xen and VMWare to Manage Virtual Machines:
A New User's Saga

John Porter, VCR-LTER

When is a computer not a computer, but really a bunch of computers? Answer: when it's running virtualization software. Each "virtual machine" runs its own operating system and can be rebooted independently of the other virtual machines running on the same host. I had several reasons for wanting to explore virtual machines.

1. **Allowing Specialization of Function**: In the past I've run a Unix-based server that combined web, database, image display, analysis and email functions. However, sometimes those functions get in each other's way. For example, one tool may require specific options be enabled in PHP, but those options may conflict with those needed for another tool. Splitting functions across machines eliminates those conflicts, since each machine can be customized to meet its specific functions.

2. **Security and Recovery**: As we all know, each of our Internet-connected machines are under near constant attack by hackers, and comment spammers. Although vigilance to maintaining updates, firewalls etc. can help, there are still times when it's desirable to be able to quickly "roll back" a machine to a previous (un-infected) state. One characteristic of virtual machines is that they are easily copied – allowing a backup copy of a ready-to-run machine to be substituted for a damaged one. There are even some cases, where it may be reasonable to "start fresh" each day (or even each hour) with a "known good" copy of a machine. If that good copy is hacked, it doesn't matter for long – since in a few hours it will be completely overwritten and erased anyway.

3. **Testing and Experimentation**: Making changes on-the-fly to a production server is always a risky proposition – something can break, perhaps unnoticed, that will negatively impact users. At the same time, maintaining a separate "development" machine in parallel is time consuming as each update in the production machine needs to be replicated in the development machine. Virtual machines make it relatively easy to create a copy of the production machine, make proposed changes, then if they are successful, substitute the development machine for the production machine.

With those ideas in mind, I've been experimenting with two free packages for creating and managing virtual machines: Xen and VMWare-server. Xen is an open-source system whereas VMWare-server is commercial software for which licenses are free. At the coarse scale, both systems take a similar approach. Each virtual machine occupies a block of dedicated disk space that appears to virtual machines as if it is a physical disk, and partition memory into dedicated chunks for each virtual machine.   Both also have a variety of ways of handling networking, such as "bridging" connections so that each virtual machine thinks it has its own Ethernet card on the main network, or "virtual networks" where the host takes on the role of DHCP server and provides network address translation (NAT) to each of the virtual hosts, allowing them to share the single public address of the host.

Given my preference for open source software, I started my experimenting with Xen (http://www.xen.org/). My host computer was running the Fedora 8 Linux variant (although I also tried it with Ubuntu Linux, with similar results) for tests of both virtualization packages. My experience with

Xen was mixed. The main documentation on installation runs along the lines of "Install the software, then edit the example configuration file to create a virtual machine." However, the example file is less than clear regarding many of the options, and truth be told, seemed to fail a lot, even when selecting default options. Fortunately, there is lots of other information on the web about how to configure Xen. The most potentially helpful was the **virt-manager** tool that provides an interface for setting up and managing virtual machines under Xen in a Fedora environment. With it, I had my first success at setting up a working virtual machine. However, many of the installation options did not work as they should. CD-drives appeared and disappeared from the options, networks worked, then didn't again. I even had a case where I was able to install a machine from the CD drive, but then once the Xen software updated through the automated Linux updates for Fedora 8, the CD drive disappeared again! Also interesting was **virsh**, a more generic virtual interface management tool that could convert configuration files to and from xml.

However, despite my best efforts, I ran out of the time I could allocate to testing Xen before I really achieved satisfactory level of functionality. This is not to say that it can't be done, just that it takes more than 3 or 4 days of concentrated effort!

I then tested VMWare-server (http://www.vmware.com). Installation had only one glitch. Fedora Linux is not one of the operating systems for which a pre-built "monitor" program exists, so it needed to be compiled as part of the configuration. Needed was a "vmware-any-any" patch (available on the VMWare FTP site) that allowed the compilation to succeed under Fedora. The **vmware** interface was similar in format to the virt-manager program used with Xen, but in my experience generally worked a lot better. In about three hours I was able to have my first virtual Linux machine up and running. Subsequently, I was able to add additional Linux and Windows XP virtual machines with little additional experimentation. I also experimented with manually "cloning" virtual machines (automatic cloning is a feature of the $189 vmware-workstation software), by copying the files containing the virtual machine to a new directory, then temporarily shutting down the original (to avoid an IP address conflict), and editing the clone to give it a new IP address and name. I installed the current production version of vmware-server (1.05), but also tried out vmware-server 2.0 beta 3, which uses a web interface instead of the standard monitor. That installation was successful, but I kept running into authentication problems when entering the monitor web site, so I went back to using the production version.

In the end, my sense is that Xen is the more versatile (and potentially powerful) product, but that it really wasn't quite ready for my use. VMWare offered slightly more limited options, but was more robust in terms of setting up working virtual machines. Fortunately both these tools are still undergoing active development, so with luck, they will both be strong contenders for use at LTER sites in the near future.

Making Sense of Collaboration Software

James W Brunt, LTER Network Office

There is much confusion about the variety and functionality of software tools in the category of collaboration software. This is really not surprising because I have found no product or class of products in recent years more hyped and loaded with jargon than that of collaboration software. The very words have a different connotation to every user, developer, and marketer. I'll attempt here to tease out the important points and make some distinctions between different classes of software. Collaboration software is commonly referred to as category of software under the general heading of "groupware". Groupware can be divided into four categories depending on the level of interaction —communication, conferencing, collaboration, and coordination.

Communication can be thought of as asynchronous interchange of information. Email, web posting, wikis, and webcasts fall into this category. **Conferencing** refers to synchronous communication between 2 or more individuals. Phone calls, Chat sessions, video teleconferencing and web conferencing systems are examples. **Collaboration** refers to interactive work toward a shared goal. Shared whiteboard applications and shared document editing are examples of this. **Coordination** refers to complex interdependent work toward a shared goal. Using project management software to develop software on timeline is an example of this.

Communication

Electronic tools for asynchronous communication of messages, files, data, documents, presentations, etc. between people and hence facilitate the sharing of information. Examples include:

- e-mail
- internet forums (also known as message boards or discussion boards) — a virtual discussion platform to facilitate and manage online text messages
- Web publishing, wikis, extranet systems, and intranet systems — collect, organize, manage and publish information for use in collaboration. Knowledge management and content management (CMS) systems fit into this category.

Conferencing

Electronic conferencing tools facilitate the synchronous sharing of information in an interactive way. Examples include:

- online chat — a virtual discussion platform to facilitate and manage real-time text messages
- audio/video teleconferencing — telephones, voip(voice over IP), h.323 (group conferencing specification), allow users to interact through talking and seeing

- data conferencing — networked PCs share applications, like whiteboards that each user can modify
- desktop sharing — users can access a shared document or application from their respective computers simultaneously in real time

Collaboration

- shared calendars — schedule events and automatically notify and remind group members
- file sharing software like briefcases, webDAV (web-based distributed authoring and versioning), and similar components in many CMS.
- shared document editing – advance forms like Google Docs and less advanced forms like Wikis.
- portals to domain specific applications

Coordination

Collaborative management tools facilitate and manage group activities. Examples include:

- project management systems — schedule, track, and chart the steps in a project as it is being completed
- knowledge management systems — collect, organize, manage, and share various forms of information; this might be done in extranets or intranets but must have project specific organization to be of use in this category.
- social software systems — organize social relations of groups – includes voting and consensus building applications.

Software that supports these interactions may be web-based or desktop client-based and maybe modular or integrated into more complex "collaboration" suites. Collaboration software suites tend to mix and match various components into a portal or desktop product for marketing. It's possible to make some generalizations about these by looking at the supported components. Most common of these, marketed as groupware or collaboration suites, include email, calendaring (includes task lists), and file sharing. In general there are no hard and fast rules about "collaboration" software. Flexible content management systems can be configured to create similar shared environments. The conventions expressed above are a product of the homogenization of available knowledge and personal experience with mature and developing products. Figure 1 compares the categorical names these software packages are marketed under with the more explicit collaboration components that they include. An analysis of this type is necessary to pinpoint the required functionality of the products being reviewed.

Software Marketing and Packaging Type	email	calendar	chat
Groupware	Yes	Yes	Opt
Web Conferencing	No	No	Yes
Collaboration	Opt	Yes	Opt
Project Management	No	Yes	No
Content Management	Opt	Yes	Opt

Figure 1. A generalized matrix of names under which groupware is marketed and the included collaboration components.

Conclusions

Collaboration components needed by the ecologists to support cross-site synthesis and collaboration include many of the functions discussed here but because of the distributed nature of the scientists and institutions involved it is unlikely that a product that brings all these things together in one package will be acceptable. For example, all scientists in the network have their own email provider – it is unlikely that they would be persuaded to use a common email framework. A similar example could be made for calendaring although there will be a necessity for project calendaring. So collaboration software needs to be modular, not dependent on one package, like email, for the whole to function. The ideal system would be web-based and/or cross-platform and support adopted standards (e.g., LDAP and H323).

Big Science and Local Meetings

Karen Baker[1] and Sabine Grabner[2]

[1] PAL-LTER and CCE-LTER   [2] MCR-LTER, Austria UAppSci

Arrangements that enhance collaboration are critical to enhancing the geographically distributed work environment of LTER information managers. Given a data environment filled with short-term deadlines in terms of both field data acquisition as well as digital systems providing data access and networked data sharing, long-term planning through new organizational mechanisms may ease the burden of planning for information exchange.

The annual LTER Information Management Committee (IMC) meeting provides a central, stable forum for our well-established information management community-of-practice. This meeting holds importance at an organizational scale. Are there additional ways to enhance community communication and coordination that may ultimately feed back to enhance the LTER IMC meeting itself? The EIMC meeting to be held in conjunction with the IMC Meeting this year is one type of new event, a reaching out and a scaling up that can broaden information management discussions. Supporting small locally organized meetings represent a type that scales down in order to network. At the implementation level, goals are frequently more easily identified or defined initially in smaller groups. Contemporary communication environments such as VTC are powerful tools but for ambitious projects these mechanisms are frequently most useful after initial personal contact and project layout.

We are exploring self-organized local meetings as an additional venue for information exchange. Note, the suggestion is for a 'local meeting' rather than a 'small meeting'. This is deliberate, namely, to avoid connoting that small activities occur at such meetings. This serves as a reminder that small-scale meetings do address big science and have ramifications for the larger community. It is only the number of initial attendees that is small in order to facilitate low overhead, informal meetings. Informal scheduling allows for unexpected turns of events, particular types of emergent synergies, and more directed conversations rather than large-scale discussions with their myriad of overlapping contexts and trajectories.

The Ocean Informatics team (PAL and CCE) routinely schedule events for 'pass-through' visitors who have formal meetings scheduled at UCSD but who coordinate a follow-on meeting with our Ocean Informatics team. Ocean Informatics events have included visits from librarians with interests in E-science and institutional repositories, an industry web manager with experience with content management systems and an archivist with insight into long-term artifact preservation, as well as information management colleagues.

In a recent meeting of participants involving PAL, CCE, and MCR, we exchanged information in particular about data system organization and data turbines. We learned from each other how we manage and publish data and metadata at our respective sites as well as about how our approaches differ. The DataZoo information system publishes data as standard packages as defined by the data provider or as temporal and/or spatial subsets, accessible through intuitive web based user interfaces. Data are available in tabular or graphic form. It was not too surprising

that the overall approach of MCR was comparable. MCR seemed on a similar trajectory given the related research topics that frame the site work. Together we began to analyze those common needs and shared user requirements among our marine and coastal sites as opposed to across the full spectrum of LTER sites. It was enjoyable to work with an underspecified agenda that allowed for unplanned, difficult-to-account-for tacit and implicit information about site and network arrangements, plans, and history .

Last year, a conversation with an LTER community advocate prompted some thinking about the value of creating an organizational arrangement in the form of a funding mechanism such that local meetings did not wait for the serendipitousness of an unrelated meeting being scheduled. We thought it valuable to transform Ocean Informatics events into a recognized and supported part of an information infrastructure. To try out this approach, support for a local meeting was included in this year's - almost regular but not taken for granted – PAL and CCE annual site supplements. Travel support was included so that a colleague could travel to San Diego for a brief stay. Site LTER co-PI's were supportive in that the funds were designated for an information management event rather than for the myriad of other site needs. A budget, however small, becomes a place marker, a seed for proactive planning and a recognition of the value of these small but formal exchanges. Having a budget specifically designated for a local information management activity opens up possibilities. We are able to plan events to match the timing of our local needs whether to investigate common approaches and themes, to share targeted applications and techniques, or to explore issues and joint articulation work.

So expect an update next year on our view of the shift from reactive planning to strategic meeting planning. Like the saying goes 'good things come in small packages'; we would add that 'large ramifications may come from local meetings'.

Preservation Metadata: Another Chapter in the Metadata Story

Lynn Yarmey, CCE-LTER & PAL-LTER

The LTER Information Management community has devoted a great deal of time and attention to building relationships and communicating with local scientists in order to meet current site metadata needs. It is also valuable to periodically review our practices and documentation with respect to future user needs. At CCE and Palmer LTER, our iTeam's attention has thus far focused primarily on implementing an efficient and stable system to fully represent the environmental data of each site. From data acquisition practices to data processing procedures, our extended EML metadata schema uses a suite of qualifiers to describe the complex data collected and analyzed by local scientists. This description captured through metadata serves the immediate exchange and access needs of our local community.

In considering the long-term however, needs change as local 'use' shifts to 'reuse' by those outside the local community and data are moved into other long-term repositories.   As data get further away from their source both in time and location, new data representation and maintenance needs arise. From a broad data archival perspective, description of data handling at the local repository level is just as important as the initial description of the original data acquisition.   This type of information is called 'preservation metadata' (as opposed to 'descriptive', 'technical', 'administrative' or 'discovery' metadata) within the library science community; preservation metadata is defined as information needed to preserve digital objects. This metadata describes the data as an object in any form (e.g. a file or a database), and includes such information as the date of upload to a system, who completed the uploading, what format and content changes occurred either at the time of upload or subsequently, who is responsible for the changes and so forth. Preservation metadata maintains the authenticity and integrity of the data over the long-term through documenting changes occurring subsequent to the original data acquisition.

As with any type of metadata, standards provide helpful prompts that stimulate community thinking about data exchange.   In 2002, the library community formed a working group (sponsored by the Online Computer Library Center and the Research Library Group) to discuss a preservation metadata framework. This process resulted in a metadata model and exchange standard called PREMIS, the Preservation Metadata: Implementation Strategies (http://www.oclc.org/research/projects/pmwg/). I recently attended a tutorial on PREMIS offered through the University of California San Diego libraries. While the main focus of PREMIS is on digital documents, there are efforts underway to map scientific data into the PREMIS structure. We can view PREMIS as a container to bring the data and their EML metadata descriptions together with information about a site's data management. Metadata about such topics as the data handling processes (Has the data been changed? Was the original data file name changed?), file format (Was the original file a .dat, .txt, .xls, .mat?   What software created the data file? What software is needed to use or view the data?), related datasets or websites, versioning information and many other elements are included in PREMIS.

When considering a new procedure or practice, the implementation process may be seen as including the work of design, development, deployment as well as enactment (Baker and Millerand, 2007). Enactment, or the initiation of use within a community, is frequently problematic. As was the case with EML implementation, the enactment phase of PREMIS is not currently supported and buy-in is slow, hindered by few available tools, crosswalks and resources.   At CCE and Palmer, we are promoting readiness for data preservation by keeping up with developments in the library and archival realms.   As LTER Information Managers, we hold a key piece of the data stewardship role, namely for supporting local data use while simultaneously bridging to reuse communities as well.   Preservation metadata initiatives provide the LTER community an opportunity to better understand the full context of data stewardship.

References:

Baker, K.S. and Millerand, F., 2007. Articulation work supporting information infrastructure design: coordination, categorization, and assessment in practice. Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS'07), January 03-06, 2007, IEEE Computer Society, Washington, DC, 242a.

Data Quality: Yet Another Chapter in the Metadata Story

Lynn Yarmey, CCE-LTER & PAL-LTER

As our LTER site datasets flood in from the field, we Information Managers, as the 'humans-in-the-loop,' remain continually challenged by the basic yet broad issue of standardization for data integration. Equally important for integration efforts as attribute-level metadata is standardization of data quality assurance (QA) and control (QC), necessary as similar datasets are brought together at sites and across the community. While data quality has been coming into its own as an issue independent of, although related to, metadata and data access, it remains largely unexplored and underdeveloped.

At an international level, quality standardization issues have been discussed for years by organizations like the International Organization for Standards (ISO) and the Organization for Economic Cooperation and Development (OECD). Recently, the Digital Library movement has begun development of sensor network data assessment tools (Wallis et al.). However, there remain many unique data quality 'quirks' associated with dynamic, local science that remain undocumented. To address these, a few grassroots communities have formed with the intent of creating community best practices and data quality standards, e.g. the LTER IM Data Quality Working Group.

One more formal example of such an effort is Quality Assurance of Real-Time Oceanographic Data (QARTOD, http://qartod.org), which has been organized and funded by NOAA. QARTOD is a cross-agency effort driven to address issues of data quality description and standardization in the context of large-scale physical oceanography with, however, secure grounding in a local, hands-on perspective. Bringing together instrument manufacturers, data collectors, metadata specialists, funding agencies and data center representatives, QARTOD looks at the quality assurance and quality control endeavor from a variety of angles and perspectives. A QARTOD meeting report states: 'One of the primary challenges facing the oceanographic community will be the fast and accurate assessment of the quality of data streaming from the [Integrated Ocean Observing System] IOOS partner systems. Operational data aggregation and assembly from distributed data sources will be essential to the ability to adequately describe and predict the physical, chemical and biological state of the coastal ocean. These activities demand a trustworthy and consistent quality description for every observation distributed as part of IOOS.'
(http://nautilus.baruch.sc.edu/twiki/pub/Main/WebHome/QARTOD2006_v9.pdf)

QARTOD is divided into measurement groups where initial active groups include CTD, in situ currents, waves and dissolved oxygen. Each group uses the creation of a common system of data flags as a mechanism for standardizing practices across disparate instrument types, manufacturers, sample analysis methods, etc. Meetings began in 2003, followed by two in 2005 and one in 2006. No plans have been made yet for the next QARTOD meeting, the community is awaiting funding decisions.

As the LTER sites continue to work towards improving data description and interoperability, the issue of data quality is likely to become more recognized as central to

integration. Community efforts like QARTOD provide a valuable forum for needed discussion, providing lessons learned and a realistic groundwork for others to build upon.

References:

Wallis, J.C., Borgman, C.L., Mayernik, M.S., Pepe, A. (in press). Know thy sensor: Trust, data quality, and data integrity in scientific digital libraries. European Conference on Research and Advanced Technology for Digital Libraries 2007, Budapest, Hungary.
(http://www.ecdl2007.org/) (see also
http://polaris.gseis.ucla.edu/cborgman/pubs/wallisetal_2007.pdf)

Specifik, a Taxonomic Reference Tool to Support Synthesis of Vegetation Data

Ken Ramsey[1] and Nicole Kaplan[2]
[1]Jornada Experimental Range LTER, [2]Shortgrass Steppe LTER

Ecologists who study vegetation dynamics in a specific ecosystem have their own set of codes for identifying and collecting plant species level data. These site specific codes are important to local researchers and create a vocabulary used by site researchers collecting species data in the field, conducting analysis, or communicating with other researchers, students, or technicians at the site. These codes are site specific and remain mostly unchanged over time as taxonomists re-classify or re-name individual species. The use of these local species codes creates an obstacle to the integration of vegetation data from multiple sites. Resolving species code differences across sites can be a significant challenge to performing synthetic analysis, especially as the number of participating sites and species of interest increase.

The Grasslands Data Integration project (GDI) found that relating participating site's plant species codes with a common taxonomic database made integration of species annual biomass observations from multiple sites significantly easier than resolving differences in site species codes. We found that if site plant species tables are standardized using common species codes, this obstacle to integration is overcome and the site's capability to integrate vegetation data with other sites or projects are significantly enhanced as many sites use their plant species codes within multiple research studies. The USDA Plants Database (PLANTS) was chosen to provide this common reference of plant species codes for GDI. However, there was no easy way to match site species codes other than exploring the PLANTS website at http://plants.usda.gov/.

Juli Mallett of the Canopy Project (http://canopy.evergreen.edu/) developed a tool to allow information managers or researchers to correlate each species within a site's species table to the PLANTS species codes. The tool, Specifik, creates a comma-separated value file (CSV) based on the original species table with the addition of a new field containing the PLANTS species code. The remainder of this article describes how to use the Specifik tool.

Procedure for using Specifik:
1. The first step in using the Specifik tool is to convert your site's plant species list to CVS format.
2. Once your site's species list is converted, you will need to go to the Specifik website (http://alala.evergreen.edu/~mallettj/specifik/). Under the section labeled 'Convert a new species table', enter a name for your new species table. (e.g. My New Species List)
3. Next, you will need to upload your plants species list (CSV) to Specifik using the Browse and Upload buttons
4. You will need to inform Specifik which columns in your species list correspond to your species codes and scientific name or genus. Note that if your species list contains a column containing the full taxonomic name, you should use this field instead of a genus field as it will make the association to the PLANTS species code easier as that column will be displayed before other fields within your species list. This will minimize the amount of horizontal scrolling when matching records in your species list to PLANTS

codes. If your site parses the genus and species into separate columns, you will need to determine whether it is worth creating a concatenated field by experimenting with Specifik using your currently formatted species table.

5. You will now start matching the individual records in your species list to PLANTS species codes. You can find matching species in the 'Possible matches from PLANTS' section.
    a. Hide help/Show help button: allows the user to hide/unhide the help text. By hiding the help text, it minimizes the amount of vertical scrolling
    b. Skip for now button: allows the user to skip the current record for now, once you have finished matching the last record in your species list, the records that you have skipped will then be cycled through again
    c. Process manually button: this allows the user to permanently skip the current record. No PLANTS code will be entered in the corresponding record in the new species list. The code will need to be entered manually into the new CSV file after the matching process in Specifik is completed. This allows the new plant species table (CSV) to be created without having to resolve all site specific species coding issues encountered.
    d. Use the selected code button: this button allows the user to match the PLANTS code selected to the current record
6. Once you have finished matching your species list you will either be prompted to save the new CSV file or you may need to use 'save as' to save the new CVS file containing your new species list. This depends on the browser.

Notes on using Specifik:
- You can select the Exit button at any time to quit the Specifik tool. You can then come back to the Specifik website later and start from where you left off!
- Many of your unknown species codes can be matched using the 'Plant codes for unknown species' section. (e.g. SOIL site code matches to 2BARE code)
- When an exact match is not found for the species or subspecies, it is recommended to match the record to the genus or species level instead.

Please send any feedback on Specifik to: mallettj@evergreen.edu

Using Keyhole Markup Language for Geographical Data Browsers

John Porter, Virginia Coast Reserve Long-Term Ecological Research

The last several years have seen a minor revolution in the online display of geographical information. The advent of Google Earth in 2005 provided an easy-to-use, flexible and attractive way of viewing land cover and topography for the entire planet. With it came Keyhole Markup Language (KML), which allows users to add data to the background provided by Google Earth and the browser-accessible Google Maps.

Before getting into some of the technical details surrounding KML it may be worthwhile to look at a sample use of KML to support data browsing. At the Virginia Coast Reserve Long-Term Ecological Research project, a number of researchers had expressed interest in being able to browse for data based on location, rather than keywords. Using KML we were able to set up a display that not only showed locations where data was collected, but also indicated which locations had more kinds of data (Figure 1, http://www.vcrlter.virginia.edu/data/vcrdata.kmz).
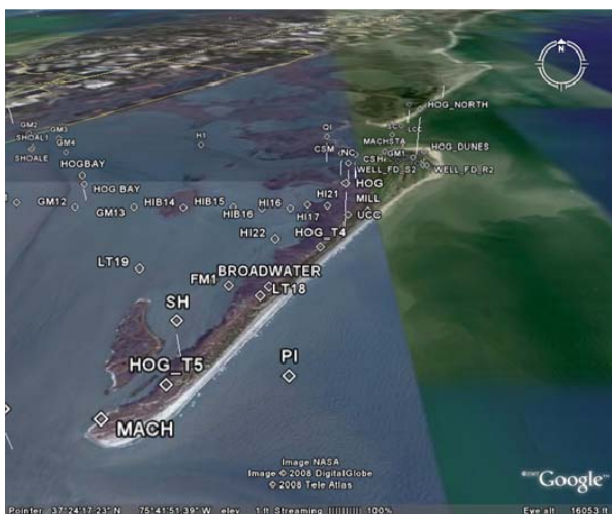

Figure 1: Locations associated with data sets on Hog Island and Hog Island Bay. The more data sets associated with a location, the higher the point is "extruded" from the background.

Developing the map display was relatively easy – we just needed to create a KML file that included the coordinates of each location, along with the number of datasets at that location. In our case, we were able to take the coordinates out of our metadatabase (or individual Ecological Markup Language documents) and use that data to build the KML file.

The technical aspects of preparing KML documents are covered in detail at the web site:
http://code.google.com/apis/kml/documentation/
which includes tutorials, manuals, and most helpfully – samples! However, here is a quick overview. First, KML documents are standard text files that use eXtensible Markup Language (XML) to provide structure. However, since text files are very inefficient regarding file size, especially for complex geographical datasets, KML documents are usually compressed using the standard ZIP algorithms into a .KMZ file. If you unzip a .KMZ file, you'll be able to see the KML document. This is very handy, similar to the "view source" capabilities of web browsers, because it allows you to view a sample application, then view the coding to see precisely how it was accomplished. KML provides several levels of organizational structure to help organize displays, including documents and folders. A <LookAt>

element provides the viewpoint from which the data should be viewed, defining latitude,

```
<?xml version="1.0" encoding="UTF-8"?>
<kml xmlns="http://earth.google.com/kml/2.1">
   <Placemark>
    <name>VCRLTER</name>
    <description>The location of the <![CDATA[<a
href="http://vcr.lternet.edu">Virginia Coast Reserve
Long-Term Ecological Research Project.</a>]]>
</description>
    <Point>
       <coordinates>-75.9252,37.2873,0</coordinates>
    </Point>
</Placemark>
</kml>
```

Figure 2: A simple KML file for a single placemark, including a hyperlink in the description element. View angle, location and icon type are not set, so defaults are used.

longitude, altitude and viewing angle for the viewer at startup. Data points are displayed using a <Placemark> element that in turn contains <name>, <description> and <Point> elements. <Point> elements then contain a <coordinates> element that provides the latitude, longitude and elevation for the point. The <description> tag can include XML <CDATA> elements that incorporate

Hypertext Markup Language (HTML), thus allowing pop-up text to include web links. Additional elements support line and polygon features, as well as points. Figure 2 contains the full KML coding for a document containing a single placemark, using defaults for viewing location, and icon style.

In addition to geographical data, individual features may be tagged with temporal coordinates (either single dates, or date ranges). A date "slider" allows users to scroll through time, selecting either a specific date or a range of dates. We have used this feature in a second application that allows users to view datasets based on both date and time (Figure 3 and
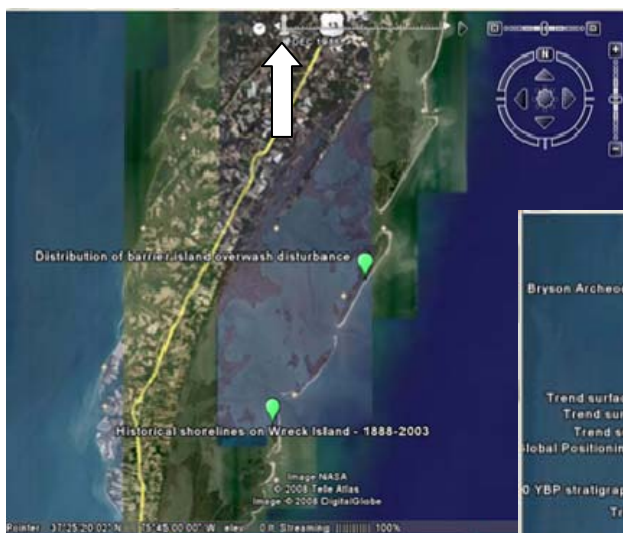

Figure 3: Data predating the inception of the VCR/LTER in 1986 (left) and during the period April 1999 through October 2002 (below). The arrows point to the time slider

http://www.vcrlter.virginia.edu/data/vcrtime.kmz). Both of the applications use data that is readily available in Ecological Metadata Language documents and maps for individual datasets can be produced using simple XML stylesheet transformations.

Finally, KML and Google Earth can also be used to capture geographical data, as well as distribute it. A toolbar allows users to add new placemarks, including lines and polygons, to a Google Earth display. A "save as" option allows users to save this data as a .KMZ file. If the user then sends you the KMZ file, you can easily extract the internal KML file, which contains the coordinates in clear text. However, a cautionary note is that although, in general, the geographical coordinates provided are not terribly inaccurate, they vary in accuracy depending on where they are. Topographically diverse areas usually experience more problems, so it is a good idea to confirm its accuracy in your research areas by identifying known points on the ground and using GPS or maps to confirm their location. Unlike most analysis-oriented GIS systems, KML provides relatively little structure for non-spatial data. Whereas ArcGIS shapefiles include both non-spatial attribute data and location data, KML files typically contain only the spatial and formatting information, plus <description> elements that allow unstructured, or arbitrarily structured text. The new <ExtendedData> element in the KML 2.2 beta holds some promise for adding some database functionality in the future.

KML continues to evolve. Version 2.2 is now in beta testing. ESRI GIS products now can output and display KML documents, so any vector data stored in an ESRI data structure can readily displayed in Google Earth. KML documents can also be displayed in a standard web browser in Google Maps, by putting the web address for the KML document in the "search maps" box. Additionally, there is a growing crop of third party software products (e.g., KMLCSV, kml2shp, Shape2Earth, Spreadsheet Mapper and a series of tools to link GPS units with KML data) that provide additional ways to use KML. Several other LTER sites also use KML. The Georgia Coastal Ecosystems LTER provides some geographical data in both ESRI Shapefile and KML forms, the Palmer and California Currents LTER sites use Google Maps and the Taiwan Ecological Research Network have developed EML to Google Earth and Google Maps converters. However, given the growing power and popularity of this technology, it is very likely that many more LTER projects will be using KML in the future!

EML 2.1.0 to be Released Soon

Margaret O'Brien[1] and Matt Jones[2]

[1] Santa Barbara Coastal LTER, [2] National Center for Ecological Analysis and Synthesis

EML is an open source, community-developed metadata standard for describing ecological data. Since 2004, it has surfaced as the *de facto* standard for international ecological groups including field stations (e.g. OBFS, LTER) and publishers (ESA, 2006), and will be integrated into the NEON observational network data management plan (NEON, 2006). Within the LTER, adoption of the EML standard has been essential to building network synthesis projects such as EcoTrends. Additionally, the LTER Network and the Genomics Standards Consortium have begun exploring a partnership to create an informational framework in which genomics datasets can be linked to environmental observations using EML (San Gil et al, in press). An API was developed in 2007 to parse EML content, download and store data entities in a relational database, and its content can also be parsed by scientific workflow software (e.g., Kelpler). The increasingly widespread adoption and application of EML is a testimonial to its comprehensive descriptions and ease of machine processing.

The EML standard is written in XML Schema Description Language, and maintained by a group of volunteers who donate their time and experience. The specification has had two major revisions (currently at version 2.0.1), and the development group is currently preparing the next release, EML 2.1. The first release of this series addresses several small bugs, and more extensive features are being considered for later releases. Because versions in the each series must be backward compatible, impacts must be carefully considered as features are developed. The user community will notice very little difference between schemas EML 2.0.0/1 and EML 2.1.0. However, users should be aware that instance documents (e.g., datasets or citations) will not be interchangeable. EML 2.0.0 and 2.0.1 documents will still be accepted.

Highlights of the EML 2.1.0 Release:

Although changes to the schema are minor, incompatibility with 2.0 instance documents necessitated advancing the version number to 2.1. In 2004, (the year EML 2 was introduced), some parts of complex schemas were left unchecked by the validation software of the day. Today's software has advanced significantly, and schema checking is more complete. This increased scrutiny of complex schemas has brought to light some EML components which require updates to remain compliant with XML community standards.

1. <additionalMetadata>: This section is designed to contain metadata which describe the resource further, but which are not appropriate in other parts of the EML document. One common use of this section is the description of units not included in the list shipped with EML (i.e., <customUnit>s). In the 2.0 series, the content model for additionalMetadata was composed of an optional element <describes> along with any well-formed xml fragment (e.g. <unitList>). In 2.1, the xml fragment must be enclosed by a <metadata> element.

2. Several items were added to the list of standardUnits, and small corrections made to descriptions of unitTypes. Although there are some inconsistencies in spelling and form among standardUnits, no changes were made to these, so that the impact on existing documents would be minimal. (see below).

3. In geographicCoverage, the <gRing> element was wrongly typed, making this element unusable. This error has been corrected.

4. The literature schema now accommodates articles which have been accepted by a journal, but are still in press. The elements pertaining to the journal's volume and page range are now optional.

5. Several elements of txt:TypeText required additional definition. These changes affect only documentation of the schema itself, and do not affect instance documents.

Other Planned Features:

Several users have requested that additional text structure be allowed in certain fields to accommodate formatted text which may lose meaning when expressed in simple strings (e.g., species binomials, chemical notation). Formatting should not be applied to fields which are designed to be rigorously machine parsed (e.g. taxonomy), or whose limitations would be better addressed by further development. However, formatting may be appropriate for text that is inherently presentation oriented (e.g. title). EML developers will review all candidates and these changes will be backward compatible within the 2.1 series.

EML's Parser will be updated to check for full schema validity. According to W3C recommendations, some parts of a schema are allowed to be treated "laxly." However most modern parsers now rigorously check all elements. If EML is to be extended in the future, its parser must be capable of fully checking any new modules. Since this parser is also used by other applications (e.g., Metacat and Morpho), there are multiple considerations to be dealt with to avoid triggering problems for content providers. This is planned to be completed for v2.1.0.

Some issues with an included unit dictionary remain and will be addressed in a future release that will not be backward compatible. Certainly the lessons learned from the LTER Network's Units Dictionary Project will be considered. The outcome may also serve as a model for handling other external structures in EML.

The EML community has four years of feedback on the schema's use and application which continues to be vital to its growth. EML 2.1 is the next step in the process by which the schema can be further extended for sophisticated annotation and searching mechanisms that are becoming available for scientific metadata, including ecology (e.g., Madin et al 2008, see Good Reads, below). EML can be downloaded from the KNB website at http://knb.ecoinformatics.org/software/eml. The site also includes documentation, general development information, and links to the database of bugs and planned features.

References:

ESA, 2006. Data Registries Workshop Report. (http://www.esa.org/science_resources/
   DocumentFiles/DataRegistry_WorkshopReportFinal.pdf)

Madin, J, S Bowers, M Schildhauer, M Jones, 2008. Advancing ecological research with
   ontologies. Trends in Ecology and Evolution, 23:159-168, doi:10.1016/tree.2007.11.007

NEON, 2006. Networking and informatics baseline design for the National Ecological
   Observatory Network. http://www.neoninc.org/documents/NIBD_2006Oct23.pdf

San Gil, I, W. Sheldon, and many others, in press. Defining linkages between the GSC and
   NSF's LTER program: How the Ecological Metadata Language (EML) relates to GCDML
   and other outcomes. OMICS: A Journal of Integrative Biology

A closer look at the NBII metadata clearinghouse

Inigo San Gil[1] and Giri Palanisamy[2]

[1] LTER Network Office, University of New Mexico, Albuquerque NM
[2] Oak Ridge National Laboratory, Oak Ridge, TN

The National Biological Information Infrastructure (NBII) has entered the fourth year of the LTER-NBII cooperative agreement. Four years is a good time to evaluate the outcomes of this cooperative effort, but rather than writing a valuable article about the benefits and lessons learned through this agreement, in this DataBits feature we are going to focus on the metadata server and tools that the NBII revamped recently.

The NBII co-sponsored a substantial effort to enhance some of their clearing-house services. The Oak Ridge National Laboratory (ORNL), another key NBII partner, hosts the NBII clearinghouse (http://mercury.ornl.gov/nbii) using a system called Mercury to harvest, search and retrieve metadata. Originally developed for NASA, Mercury's business model is based on a consortium of projects that share general costs and share the benefits. There are over a dozen projects supporting Mercury, and as we mentioned NBII is in this international consortium. Mercury provides a single portal to information contained in disparate data management systems, reflecting the data and metadata diversity that is managed at ORNL Distributed Active Archive Center (DAAC). Mercury is also standards based – XML, EML, FGDC, Z39.50, Dublin-Core, Darwin-Core, GCMD and ISO19115

Searching and plowing through results made easy

The Mercury enhancement consisted of changing the indexing and searching interface from a proprietary implementation to an open source counterpart that mimics the Google text indexing approach (Brin, 1998) to classify, categorize and speed up searches based on rankings. This open source Apache project, codename Lucene (Gospodnetic, 2003), is used in conjunction with Solr (Delacretaz, 2006), another Apache open source project that extends the functionality of Lucene by giving proper consideration to numeric types, dynamic fields, and unique keys. An example to clarify what this means: Solr gives the developer the ability to give special treatment to specific geo-temporal coordinates. Special information that is used in an advanced search can be treated properly using Solr, as opposed to be buried among the competing rankings given by Lucene to all the metadata content.

Figure 1 is a snapshot of the 'advanced search' interface used for the NBII Clearinghouse. Note the ability to narrow the search to the LTER data catalog (bottom right quadrant). The ORNL-DAAC hosts a wealth of national and international metadata resources (http://mercury.ornl.gov/ornldaac); however, LTER ranks highest in unique volume of contributions. Also, there is a Google Maps mash-up that allows users to draw a bounding box or enter a location, very much like in LNO's metacat advanced search

interface. The browse function (top right) offers a drill down approach to data search – there is no left menu, this seems to be the right trend on efficient web layout these days.

Figure 1 – a snapshot of the Mercury advance search interface (NBII)

The new technologies are based on service oriented architecture, and the Google Maps mash-up is an example of such ability. Other example are RSS feeds that can be easily created for frequently repeated searches. Mercury also provides the harvested metadata to other applications (eg., Google, NASA Global Change Master Directory , NBII Biobot).

Surviving the results page

   The results page has also been revamped. The top right offers push buttons to create an RSS feed, bookmark or an email for these results. RSS or bookmarks enable refreshing the query matches periodically without the hassle of recreating the query. The aged left menus have disappeared, lending the added web real state to the actual content, which is stacked in two sections (See Figure 2)

Figure 2 – A typical look at the query results page

Results Page Top Part – This first half of the page results offers further filtering (narrowing) of the results, with filtering choices such as "data provider, parameter (keyword), topic or project. Below filtering we find navigating and sorting controls. Here the user can navigate the results with a navigation bar and reorder the results according to selectable criteria such as time scope, source, or project. The default ranking order is the relevancy, the google-like rank assigned by the indexing process

Results Page Bottom Part – The second half of the page results offers the results; snippets of the records that match the search/browse criteria, and a link to the full metadata or associated data. The stars shown at the bottom of each record indicate the relative relevance of the matched criteria. The snippet includes the title and study date range, source provenance and excerpts from the abstract.

Styles used to display single metadata records

Finally, Mercury offers two styles to display a full metadata record. Mercury by default offers a classic, well organized reduced style at the full record and additionally, it offers what it is known as the FGDC style, which would be very familiar to those who use the ESRI tools or that have used the previous Mercury. It is plain text divided in 6 sections, with the underlying hierarchy preserved as indentation. Not a particularly appealing choice to this author.

One can argue that the most common practices in the web would dictate that the navigation bar at the bottom of the page rather than the middle. Perhaps this would free up some real state in a page whose centerpiece should be the results

Usage, practices and future plans

As the many merits, we can highlight the incredibly fast performance of the new Mercury system. After using the LTER Data Catalog based on the Metacat system, we were expecting latencies of the order of ten seconds or more for full text single and advanced searches. With Mercury we receive relevant results one order of magnitude faster, in about one second. The email, bookmark and RSS feeds capabilities are very interesting, as well as the narrowing and sorting functionalities, and it is promising that there is a plan to add further functionalities such as, thesaurus, gazeteer, openSearch, openDap support and implementation of the web service harvesting standards among other plans.

Perhaps we ought to warn about the perils of the new applications. The new Mercury has just been released, and the developers are working through some glitches that you may find. For example, restricting my advanced search to the state of New Mexico and the keyword "biodiversity" did not prevent Mercury from offering me results for studies conducted at the Kellog Biological Station (Michigan), but this may be an artifact of poor metadata or a bona fide bug in the newly released system. Other odd behaviors and kudos can be reported to the team leader, Giri Palanisamy at ORNL.

We should mention that the best feedback that this team can receive is though the use and peruse of its interface. Search activity is logged and the activity is analyzed to build better controlled vocabularies. Controlled vocabularies are used to build the browse tree searching tool as well as web services for metadata entry tools and editors. In other words, DAAC pages usage and flow analysis provide valuable feedback to enhance the DAAC human experience. The ORNL DAAC is in that regard a community driven tool, not necessarily the designer's vision of the community use. In LTER we learnt [Karasti, 2007] that valuable community tools may stall at the community engagement stage, which brings up the quest for the much needed and perhaps low-tech community process that would legitimate and elevate many stalled projects to a network wide status – and so the community challenge goes to our ToDo list.

To close this feature, a simple surf away! May be your study based on these brand new integration and synthesis enabling tools will be the next case study!

References

Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Proceedings of the Seventh International World Wide Web Conference* **v30**, Apr 1998, pp 107-117

Otis Gospodnetic, Jan 2003. Apache Leucene.
http://www.onjava.com/pub/a/onjava/2003/01/15/lucene.html .

Bertrand Delacretaz, Aug 2006. Solr – indexing XML with Lucene and Rest.
http://www.xml.com/pub/a/2006/08/09/solr-indexing-xml-with-lucene-andrest.html.

Helena Karasti, Karen S. Baker and Katharina Schleidt. 2007. Digital Data Practices and the Long Term Ecological Research Program. 3rd International Digital Curation Conference, 11-13 Dec 2007,

LTER Data Catalog Upgrades to Metacat 1.8.0

Duane Costa and Mark Servilla, LTER Network Office

Introduction

Metacat is a flexible database storage system for XML formatted documents that has served as the basis for the LTER Data Catalog since 2004. Last December, the Knowledge Network for Biocomplexity (KNB) project announced its latest release, Metacat 1.8.0, and the LTER Data Catalog was upgraded to this new version in April, 2008. This article summarizes new features of the Metacat 1.8.0 release, highlighting those that are especially pertinent to LTER users.

Performance, Performance, Performance

The overriding goal for Metacat 1.8.0 was to focus on improving the performance of search queries. The Metacat developers rolled up their sleeves, got under the hood, and found several important ways to accomplish this.

1. SQL Statement Restructuring

Metacat developers analyzed the SQL query statements that Metacat composes when matching search criteria against the actual metadata stored in the database, restructuring several of the SQL statements in a way that optimizes search performance.

2. Query Caching

Not satisfied with what they had accomplished through SQL statement restructuring, the developers went a step further and turbo-charged Metacat's search engine by adding a new mechanism to automatically cache query results. After the first time a given search is conducted, subsequent searches using identical search criteria will cause Metacat to return its results with blazing fast speed.

It's easy to envision how the new query caching mechanism can reap rewards when applied to commonly used search terms: the more common the search term, the more likely it is that it will have been entered at least once before by some other (or the same) user. However, there are some important limitations to the caching mechanism:

- Metacat has a limit on the number of search queries that are cached. By default, Metacat caches the results of up to 1000 queries, though this number can be adjusted by the local Metacat administrator.
- The cache mechanism only works with *public* users (i.e. users that have not logged in) because Metacat only caches search results for metadata documents that are readable by all users. Users that conduct a Metacat search while logged in to their KNB account can generate search results that are potentially different from those of public users in that they may contain additional metadata documents readable by the logged-in user but *not* by

public users. Thus, logged-in users will not see the benefit of caching. This limitation may be addressed in a future Metacat release.

- Whenever the Metacat database is changed by the insertion, update, or deletion of any document, the entire cache is invalidated and wiped clean. A data catalog that is frequently changing will lose its cached results just as frequently, making the likelihood of repeat searches on a given search term, even the most common of terms, much less probable. (With its daily automated harvests, the LTER Data Catalog falls into this category.)

We wanted to take Metacat 1.8.0 on a test drive to see the new caching mechanism in action. To do this, we set our browser to the LTER Data Catalog (http://metacat.lternet.edu) and entered "snake" into the search box. Next, we selected the "Search all fields (slower)" radio button to ensure that Metacat would do a thorough search of all document fields, thereby forcing Metacat to run in its slowest search mode. Finally, we clicked the "Search" button. Metacat returned its search results (60 data packages found) in about eighteen seconds—not terrible, though not terribly impressive. It was on subsequent searches of "snake" that the power of the caching mechanism became evident—Metacat returned its search results on our second, third, and fourth searches in the blink of an eye!

3. Other Performance Enhancements

An assortment of additional performance enhancements have been incorporated into the Metacat 1.8.0 release, including:

- performance optimizations specific to the PostgreSQL RDBMS;
- performance optimizations specific to the Tomcat web application server; and,
- the addition of new indices on key columns in the Metacat database.

4. Performance Summary

Metacat 1.8.0's optimized SQL statements, its new caching mechanism and other performance enhancements, as well as recent upgrades to server hardware at the LTER Network Office, make this version of the LTER Data Catalog the fastest and sleekest to date. We can utter a collective "*Adios!*" to the days of five-minute Metacat searches!

Prettier EML

The Metacat developers did a major overhaul of how Ecological Metadata Language (EML) documents are rendered in HTML. The XSLT stylesheets Metacat uses to accomplish the transformation of EML documents from XML to HTML were reorganized to:

- display citations in an optional new longer format;
- make download options more obvious to the user; and,
- display data tables using a more human-readable and machine-printable layout that involves less horizontal scrolling within web browsers.

More Supple Skin

Organizations that deploy Metacat typically specify a custom interface, or *skin,* to be used as the data catalog's presentation layer. For example, the LTER Data Catalog uses a locally developed and maintained *lter skin* which includes LTER-specific components, such as the Advanced Search form (designed by workshop participants at the 2004 Information Managers Meeting in Portland, OR).

Prior to Metacat 1.8.0, all Metacat skins were required to conform to a highly-structured IFRAME layout. (In HTML, an IFRAME element creates an inline frame that contains another document.) IFRAMEs impose certain restrictions on Metacat skin implementations that can limit their flexibility and make their development and maintenance more challenging. Put another way, IFRAME-based Metacat skins are cranky, as are the Metacat skin developers forced to use them!

Metacat 1.8.0 liberates the Metacat skin developer from conformance to the IFRAME edict. Organizations now have the option of developing their custom Metacat skins with an IFRAME-less layout, and some have already done so. The LTER Data Catalog has not yet taken advantage of this new feature; its *lter skin* is still based on the old IFRAME structure. However, it's noteworthy that future versions of the LTER Data Catalog may benefit from this new freedom and flexibility.

New URL Syntax

Metacat 1.8.0 recognizes a new, simpler syntax for direct document access. The old syntax, which is still supported, provides direct access to a document using a URL such as:

http://metacat.lternet.edu/knb/metacat?action=read&qformat=lter&docid=knb-lter-sgs.1

In this example, *docid=knb-lter-sgs.1* refers to the requested document and *qformat=lter* specifies the output style as the *lter skin*. The problem with URLs using this syntax is that many search engine web crawlers won't follow links that are laden with request parameters. With this in mind, the Metacat developers added support for a new syntax, free of request parameters, that is kinder to web crawlers (and to people, too!):

http://metacat.lternet.edu/knb/metacat/knb-lter-sgs.1/lter

Summary

Although this article has attempted to describe the most salient new features of Metacat 1.8.0, a number of other features and bug fixes were not included here. For a full description of Metacat 1.8.0 and its capabilities, download the latest Metacat release at:

http://knb.ecoinformatics.org/software/download.html#metacat

The README file contains a complete summary of all changes in the release.

We are pleased to be using Metacat 1.8.0 at the LTER Network Office as the basis of the LTER Data Catalog. We congratulate the talented cadre of Metacat developers who, through their diligent efforts, have made this important product available to the Ecoinformatics community.

Good Reads: Advancing ecological research with ontologies

Joshua Madin, Shawn Bowers, Mark Schildhauer, Matthew Jones, 2008, "Advancing ecological research with ontologies." Trends in Ecology and Evolution, 23:159-168, doi:10.1016/tree.2007.11.007

In ecology many terms are used in both metadata creation and searches which have multiple interpretations and variable meanings and contexts. This ambiguous terminology leads to redundant effort and slows scientific progress and collaboration. The relationships among data elements may be obvious to scientists, but their formal description is necessary to allow computers to deduce them. Ontologies are one mechanism for describing these relationships, and have been successful applied in some biological disciplines (e.g., microbiology). Their use facilitates data discovery, interpretation and integration.

This paper presents an overview of several ontologies which are key to ecology. It presents two basic types: domain-specific ontologies capture specialized terminology for a field of research, and framework ontologies define concepts and relationships. It also discusses other less formal methods of constraining metadata, e.g., controlled vocabularies and glossaries, and suggests uses and limitations of these. In addition to the overview, highlighted boxes which are illustrative for novices exemplify general aspects of ontology development, construction and potential errors. It is clear that the use of ontologies can enhance the descriptive power of metadata. However, community driven development and endorsement are crucial to their success. This paper can help both information managers and project scientists to understand this emerging tool.

-Margaret O'Brien, Santa Barbara Coastal LTER

Good Reads: Cyberinfrastructure Primer

A. Gold, 2007. Cyberinfrastructure, Data, and Libraries, Part 1. D-Lib Magazine 13(9/10). 13(9/10). http://www.dlib.org/dlib/september07/gold/09gold-pt2.html

A. Gold 2007. Cyberinfrastructure, Data, and Libraries, Part II. D-Lib Magazine 13(9/10). http://www.dlib.org/dlib/september07/gold/09gold-pt2.html

The author provides a comprehensive yet straight-forward introduction to scientific cyberinfrastructure for science data from a library perspective, summarizing both library infrastructure issues of today and E-Science promises for the future. Gold provides a timely overview of important elements amidst what appears as an ongoing transition in scientific research data and data practices: "To be able to exchange data, communicate it, mine it, reuse it, and review it is essential to scientific productivity, collaboration, and to discovery itself."

The reader is provided a path through important community reports that document the ongoing processes: the NSF Atkins report (2003; see Databits Spring 2005 Good Read) to the National Science Board Long-Lived Data report (2005) and finally the Association of Research Libraries (ARL) on Long Term Stewardship of Digital Data Sets. The Task Force on E-Science sponsored by the ARL has taken into consideration the changing roles of libraries and librarians in terms of data and document artifacts. Discussion ranges broadly from supercomputers and grid science to data preservation and curation. Though mention of life cycles and raw data seems to imply near-linear and well-bounded entities rather than complex concepts, the multiple dimensions of infrastructure are covered including data policies and business models. For libraries anchored by physical collections and archive traditions, considering digital data represents an entirely new aspect of their work. An important case is made for a wide range of new roles and supportive services associated with this postulated work with data.

Part 2 of this primer continues with the services offered by data library efforts today: social science, GIS, bioinformatics and archival data services. Gold makes clear that data librarianship requires not only a new framework because it differs significantly from artifact-based collections but also a re conceptualization of roles and collaborative undertakings. She wisely mentions the need to build both capacity and understanding of technology and scaling as well as of new roles. Librarians are stretching from curating physical artifacts to curating digital materials as their focus on cataloging and archiving broadens to include a new service called data management. The primer ends with suggestions for actions that librarians and library leaders may take as they venture into this new territory. This two-part primer includes a suggestion, a tap on the should, to remind the library community to "encourage conceptual dialogue regarding data and informatics efforts", that is to say, some scholarly thinking to accompany the library plunge into oceans of data.

- Karen S. Baker, CCE-LTER & PAL-LTER

Good Reads: Digital Data Practices and the Long Term Ecological Research Program

This paper was the recipient of the best peer-reviewed paper award at the Digital Curation Conference in 2007. It provides a great perspective of LTER data practices and challenges, and an enlightening discussion on the consequences of automated data processing in our pipelines.

Citation:

Karasti, H., Baker, K.S. and Schledit, K. 2007. Digital Data Practices and the Long Term Ecological Research Program. Third International Digital Curation Conference, 11-13 Dec 2007, Washington, DC, USA (http://interoperability.ucsd.edu/docs/07Karasti-Baker-Schleidt-DCC07.pdf)

Abstract:

This paper explores data curation in a Long Term Ecological Research (LTER) setting. It describes a number of salient data characteristics that are specific to the discipline and outlines some central features of the curation approach cultivated within the US LTER network. It goes on to identify recent developments within the international LTER program relating to data issues: increasing heterogeneities due to LTER networking, integration of data from additional disciplines, and new technologies in a changing digital landscape. Information management experience within LTER provides one example of the recurrent balancing inherent to the work of data curation. It highlights (1) taking into account the extended temporal horizon of data care, (2) aligning support for data, science and information infrastructure, and (3) integrating site and network level responsibilities. LTER contributes to the inquiry into how to manage the continuity of digital data and to our understanding of how to design sustainable information infrastructure.

- Inigo San Gil, LTER Network Office

USFS Experimental Forest and Range Information Manager Meeting

Don Henshaw[1] and Dave Rugg[2]

[1]Andrews Experimental Forest LTER, [2]USFS Northern Research Station

A historic first meeting of U. S. Forest Service (USFS) and cooperator information managers was held in Dallas, Texas on 5-7 February 2008. An information manager attended with a site science representative for most of the 17 USFS Experimental Forest and Range (EFR) network of sites. Concurrent with the 100-year anniversary of the EFR concept, this network now begins work in establishing a new research platform to enable study of transcontinental questions concerning effects of environmental change on ecosystems function and services. These 17 sites, which include 6 LTER sites, were selected in 2007 to organize and operate as a national research network and are a subset of nearly 80 national EFR sites. The long-term vision is to bring the other EFR sites into the network. Goals of the meeting included beginning development of an operating framework for information management and the selection of synthesis research projects and the implications for information management (IM) activities.

Discussion of an operating framework for IM to facilitate synthesis science highlighted the first meeting day. Several desired characteristics of the EFR Network are summarized here:

- a national IM coordinator position or IM team to establish network-level standards for IM and assist in their implementation, e.g. EFR Network Office;
- core data sets are compatibly measured and reported across the network;
- cross-site studies are encouraged and supported within the research framework;
- collaboration with other national research networks augment FS efforts;
- websites effectively share research results and data, and market the network;
- technology and in-person meetings effectively maintain strong communication

This EFR effort is supported through a USFS initiative, eResearch, with one goal to promote IM at EFR sites and model IM approaches after the Long-Term Ecological Research (LTER) program. The information managers met independently on the second day with special emphasis given to understanding the breadth of existing IM practices and standards currently used in both LTER and USFS. Featured talks on LTER IM standards (Don Henshaw, AND) and on USFS R&D IM standards (Dave Rugg, e-Research EFR Project Manager) described IM activities including adopted standards, network information system development, data archiving practices, and data access policies for their respective networks. Site-oriented talks allowed presentation of current site IM practices and included LTER talks by Eda Melendez (LUQ), Jason Downing (BNZ), Chris Eagar (HBR), Jim Vose (CWT), and Henshaw (AND), and several EFR (non-LTER) talks. Needs described by participants included data organization and catalog development, metadata creation, metadata standards, data search and retrieval capabilities, cyberinfrastructure capabilities, and human resources. It became clear from these site-oriented talks that the needs for the EFR sites are congruent with those of LTER sites and the EFRs could likely learn from the collective LTER experience.

The USFS R&D Data Archive, engineered by Dave Rugg, provides a central repository for USFS research data and uses the Biological Data Profile (BDP) as its metadata standard, compliant with the USFS-adopted FGDC metadata standard. Rugg has also written Metavist, a

popular interactive program for entering BDP metadata, with stylesheets to produce BDP and FGDC structured metadata. The IM meeting group favored an open approach to sharing data externally and adopted the archiving standards embodied in this e-Research R&D Data Archive project, including BDP as the metadata standard. Extensions to the data archive, such as stylesheets developed by the LTER Network Office that can transform BDP XML files into EML files and visa versa, will be added to enhance interoperability with the LTER sites. Differences in the granularity of metadata elements makes the BDP to EML transformation problematic, and best practices for use of Metavist (similar to the LTER-developed EML best practice document) will be important to improve the quality of transformation. The intent is for EFR metadata stored in the USFS data archive to be searchable through both the NBII clearinghouse system and the replicated EML-based Metacat servers used by LTER.

The meeting researchers developed six synthesis research topics and discussed these with the information managers on the third meeting day to gain an understanding of the data-related feasibility of the ideas. The EFR sites will now identify the long-term records and build a network data set catalog to facilitate these synthetic research efforts. A useful post-meeting development was the USFS data archive receiving permission to implement a data access / data use policy similar to the LTER policy in tracking who is obtaining data sets and what the planned use is. This type of policy is considered essential in promoting a culture of data sharing.

Calendar

Event: Cyberinfrastructure for Environmental Observations, Analysis, and Forecasting
Location: Boulder, Colorado, USA
Dates: May 5-7, 2008
Web Site: http://www.cyberobservatories.net/events/
Description: The Cyberobservatories initiative hosts workshops to share information on NSF and related funded Cyberinfrastructure initiatives that may provide opportunities for leveraging shared cyberinfrastructure across environmental observatory programs. A series of presentations describe technical innovations, effective management and coordination practices, and other lessons learned in the implementation of successful existing CI projects. Breakout sessions are aimed to provide opportunities for discussion and networking between discipline scientists and CI practitioners.

Event: LTER Science Council meeting
Location: Baltimore, MA, USA
Dates: May 7-8, 2008
Web Site: http://intranet.lternet.edu/meetings/
Description: LTER annual science council (SC) meeting with representatives from all sites and the network office.

Event: Scientific and Statistical Database Management
Location: Hong Kong, China
Dates: July 9-11, 2008
Web Site: http://www.ssdbm.org/
Description: This international conference brings together scientific domain experts, databases researchers, practitioners and developers for the presentation and exchange of current research on concepts, tools and techniques for scientific and statistical database applications. SSDBM provides a forum for original research contributions and practical system design, implementation and evaluation. Individual themes differ year to year with the main focus remaining on databases theory and application in the scientific and statistical fields.

Event: Ecological Society of America Meeting
Location: Milwaukee, WI, USA
Dates: August 3-8, 2008
Web Site: http://www.esa.org/milwaukee/
Description: The Ecological Society of America's (ESA) annual Meeting, held each August, draws more than 3,000 professional ecologists from around the world to participate in scientific presentations, symposia, workshops, field trips, and a trade show. This year the meeting theme is "enhancing ecological thought by linking research and education."

Event: LTER Information Management Committee Meeting
Location: University of New Mexico, Albuquerque, NM, USA
Dates: September 8, 2008 – September 9, 2008
Web Site: http://intranet.lternet.edu/im

Description: Annual LTER information management committee meeting (IMC) with representatives from all sites and the network office.

Event: Environmental Information Management Conference
Location: Albuquerque, NM, USA
Dates: September 10-11, 2008
Web Site: https://conference.ecoinformatics.org/index.php/eim/eim2008
Description: This conference is intended to bring together informatics practitioners, developers and environmental scientists interested in technologies that enable data collection, description, curation, discovery, access, integration and analysis in all disciplines of environmental research. It provides a forum to build partnerships, explore solutions to the common challenges faced by environmental observatories, and to present advances in community standards, practical system design, implementation and assessment. Share your experiences and knowledge both informally and formally with an oral presentation that will be published in the Conference Proceedings, or a poster or a demonstration of new developments. Provide input to the discussion of future meeting venues and formats.

Event: Participatory Design Conference
Location: Bloomington, Indiana, USA
Dates: September 30-October 4, 2008
Web Site: http://pdc08.informatics.indiana.edu
Description: Participatory Design (PD) is a diverse collection of principles and practices aimed at making technologies, tools, environments, businesses, and social institutions more responsive to human needs. A central tenet of PD is the direct involvement of people in the co-design of things and technologies they use. Participatory Design Conferences have been held every two years since 1990 and have formed an important venue for international discussion of the collaborative, social, and political dimensions of technology innovation and use.

Event: American Society for Information Science and Technology
Location: Columbus, Ohio, USA
Dates: October 24-29, 2008
Web Site: http://www.asis.org/conferences.html
Description: 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 what might be 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; individuals who share a common interest in improving the ways society stores, retrieves, analyzes, manages, archives and disseminates information, coming together for mutual benefit.

Event: Computer Supported Cooperative Work Conference
Location: San Diego, CA

Dates: November 8-12, 2008
Web Site: http://www.cscw2008.org
Description: The conference brings together top researchers and practitioners who are interested in both the technical and social aspects of collaboration. In recent years the conference has moved beyond traditional "work" to include the broader issues of how we play, socialize, and compete - all forms of collaborative activity that are now mediated by technologies. As more and more people in all regions of the globe are able to interact online we are rapidly moving toward a Computer Supported Cooperative World. Appropriate topic areas for CSCW include all contexts in which technology is used to mediate human activities such as communication, coordination, cooperation, competition, entertainment, education, medicine, art, and music.

Event: Digital Curation Conference
Location: Edinburgh, Scotland
Dates: December 1-3, 2008
Web Site: http://www.dcc.ac.uk/events/dcc-2008
Description: 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. The Digital Curation Conference (DCC) is held annually since 2005.

Event: International Conference on Ecological Informatics 6
Location: Cancun, Mexico
Dates: December 2-5, 2008
Web Site: https://conference.ecoinformatics.org/index.php/isei/isei6
Description: Ecological Informatics emphasizes information processing from genomes to ecosystems, meta-information concepts for ecological data management, computational ordination, clustering and forecasting of complex ecological interactions, and facilitating informed decision making for sustainable ecosystem management. The conference theme of ISEI6 will be 'Data and Software Sharing: Key to Sustainable Ecological Solutions'. The ISEI6 organisers feel committed to overcome the current fragmentation and incompatibility of ecological data and information, and promote local and global networking, sharing and exploration of ecological information by means of advanced computer technology.

Event: Hawaii International Conference on System Sciences
Location: Hawaii
Dates: January 5-8, 2009
Web Site: http://www.hicss.hawaii.edu
Description: Since 1968 the Hawaii International Conference on System Sciences (HICSS) has become a forum for the substantive interchange of ideas in all areas of information systems and technology. The objective of the conference is to provide a unique environment in which researchers and practitioners in the information, computer and system sciences can frankly exchange and discuss their research ideas, techniques and applications. To realize this objective and to facilitate lively discussion and interaction, the format is carefully structured, and the number of available registrations is limited.