ï»¿

Feature Articles

Editorials

News Bits

Good Reads

DataBits: An electronic newsletter for Information Managers ----- Fall 2006 Issue (http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/06fall/)

Past issues-- (http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/)

Featured in this issue:

This issue of Databits follows the September, 2006 LTER Information Managers' Meeting in Estes Park, Colorado. The IM meeting was held in conjunction with the LTER All Scientists' Meeting and included a number of Information Managers from around the world. In recent years, there has been an emphasis on establishing relationships among members of the International LTER community. This issue of Databits reflects this initiative with contributions from several ILTER participants. In addition to these contributions, there are a number of feature articles, news items, editorials, and suggested readings that we hope you find interesting and informative.

DataBits continues as a semi-annual electronic publication of the Long Term Ecological Research Network. It is designed to provide a timely, online resource for research information managers and to incorporate rotating co-editorship. Availability is through web browsing as well as hardcopy output. LTER mail list IMplus will receive DataBits publication notification. Others may subscribe by sending email to majordomo@lternet.edu with two lines "subscribe databits" and "end" as the message body. To communicate suggestions, articles, and/or interest in co-editing, send email to databits-ed@lternet.edu.

----- Co-editors: Brian Riordan (BNZ), John Campbell (HBR)

Feature Articles

Developing Collaborative Information Management Systems

- John Porter (VCR)

Within ecology, the tradition has been to develop stand-alone information systems, wherein each system duplicates the same basic functionality. Each information system would provide mechanisms for capturing, searching and disseminating data and metadata from and to its users. This individual-based development has been driven by project-specific perspectives and funding, wherein each research project is responsible for managing its own data. Additionally, lack of commonly used standards kept developers from developing broadly applicable tools. However, recent changes in computer technology and the development of widely-used standards for metadata create new opportunities for the development of collaborative information systems, where functionality is shared across multiple systems.

For example, imagine an information management system that provides a data catalog for locating data sets. Using the traditional approach, each of multiple information management systems would replicate that function using different software and/or approaches. However, using a collaborative approach, only one system needs to develop that functionality, and it can be exploited by a large number of other systems through â€œportalsâ€• that give the appearance of belonging to a specific system, but in fact are shared among many systems. For example, the LTER Metacat server found at: http://metacat.lternet.edu/knb/index.jsp can be transformed via a â€œskinâ€• to serve as the data catalog for the McMurdo LTER site in the Antarctic (http://metacat.lternet.edu/knb/index.jsp?qformat=mcm).

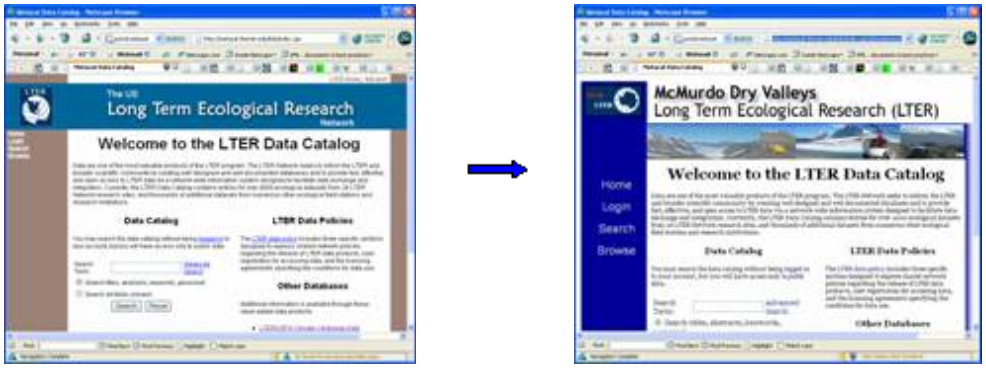


Figure 1: A "skin" transforms an LTER-wide Data Catalog to a site-specific data catalog.

The information management team of the Taiwan Ecological Research Network (TERN) has been active in developing applications that can be used collaboratively. For example, they have developed a tool for displaying research locations using â€œGoogle Maps.â€• The application is normally invoked using a web form wherein a user provides information on where to find an Ecological Metadata Language document containing the needed coordinates. However, by building a pre-filled-out form into a web page, an automatic link for showing the locations from a specific dataset can be created. The web site using the application need only provide the â€œbuttonâ€• in order to be able to use the functionality of the application.

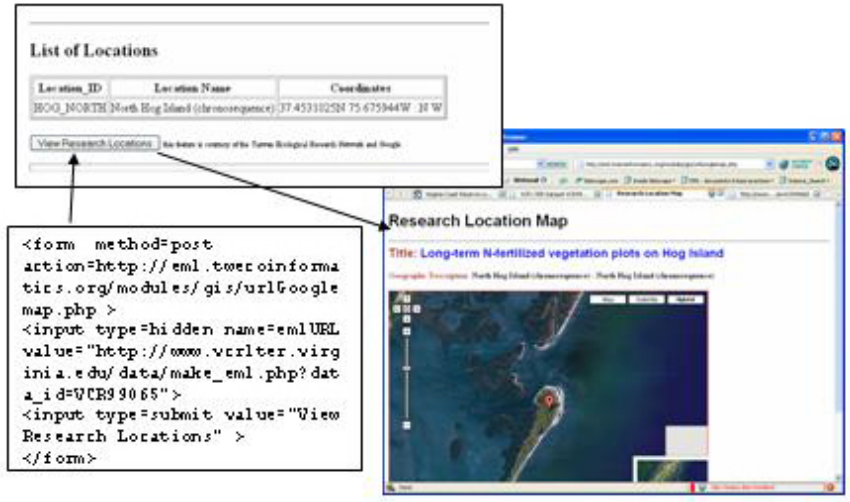


Figure 2: A web form â€œbuttonâ€• contains all the information needed to trigger the plotting of research locations for a dataset. The â€œbuttonâ€• need not reside on the web site actually doing the processing. Here the Virginia Coast Reserve LTER site provides the â€œbuttonâ€• and the underlying EML document, while the TERN site provides the application to interface to Google Maps.

In some cases it is not necessary to embed an entire form into a web page to use a collaborative tool. Any page containing a

link to: http://www.vcrlter.virginia.edu/data/eml2/getsas.php?knb_package=knb-lter-vcr.49.4 (you can substitute your own KNB Package ID) will be able to create a SAS job from an EML document. (See the Spring 2006 Databits for details on other EML tools for the R and SPSS statistical packages).

Use of collaborative applications holds tremendous promise for improving the functioning of ecological web sites. However, it also raises some ethical issues. Just as no reputable scientist would use the data of another researcher without providing credit in the form of a citation or acknowledgement, web sites using external applications need to have a way of providing credit to the site doing the actual processing. In the example above a label by the button notes that "this feature is courtesy of the Taiwan Ecological Research Network and Google." However, it is also possible to build a citation into the application itself. For example, the Google maps have built-in labels that indicate the source of the imagery. The article: "Three Challenges in Supporting Shared Workspaces" in this issue also lays out some additional challenges.

The examples above build on the very basic functionality provided by web forms. However, the increasing sophistication of tools based on web services should make collaborative information systems even more powerful. With web-form-based tools, the output is provided wholesale by the service provider. However, with web services, the product of the service is an XML file. This product file can then be transformed using XSLT to any form desired by the client system. This allows the production of customized products that can extend beyond web pages to include the transmission of data between systems.

With collaborative tools we can imagine a new era of ecological information systems. Instead of duplicating functionality already provided elsewhere, each system could concentrate on providing collaborative tools focused on specific needs. For example, one site might provide services associated with data catalogs, while another provides spatial search tools and yet another site might focus on management of bibliographic data. Each participating site has the full functionality of the suite of collaborative tools with advanced search engines, mapping and analytical tools, but only needs to provide the interface to those tools for their users.

Developing Data Ethics for a Scientific Information Society (Part 1)

- Avinash Chuntharpursat, South African Environmental Observation Network (SAEON)

With continuous advancements in Information Technology, public access to information and data has been made much easier than the past. These advancements have often resulted in swift societal change to make optimum use of the new technologies. This cascade of advance and response needs to be based on a solid ethical foundation to ensure sustainability and societal stability.

The above scenario is of particular relevance to information and data management in Long Term Ecological Research (LTER). Global climate change as a contemporary issue has exposed LTER to the public eye with the resultant demand for data potentially overwhelming scientists and information/data management systems alike. A societal data ethic should govern the way we produce, archive, extract, analyse and use scientific data.

Currently, attempts are being made with various in-house data policies and national legislature to govern the dissemination and use of data, these have their limitations. Legislature is often difficult and expensive to implement and at times unable to follow the rate of technological advancement. In-house data policies are often only relevant to the organisation that produces it and potentially mutually exclusive from the policy of another organisation hence preventing data exchange. A societal effort needs to be made to transcend these little policies to give the public greater access to data while protecting the integrity of the data and associated researchers.

At the South African Environmental Observation Network (SAEON), an ethical framework is currently being developed. Figure 1 illustrates this.

Figure 1: SAEON Information and Data Management Ethical Framework. (Note, ICT = Information and Communication Technology)

In Figure 1 the Information and Data Management System (I/DMS) is shown to consist of components and requirements. (More on components and requirements can be gleamed from a literature search.) To integrate the components and requirements there numerous ethical issues. These issues can be viewed as interactions between and within components and requirements of the I/DMS. As an example the ethical interaction between "People" and "Data Quality" could be integrity, honesty, accuracy, etc. Numerous such interactions can be derived from this framework. If anyone is interested in seeing how many they can get, please email the results to avinash@saeon.ac.za before 30 Jan 2007. (If there's enough, the results can be published in a forthcoming article, giving full credit those who derived them.)

This article was written with the aim of catalysing the process to develop a Data Ethic that is applicable to everyone. The point illustrated in figure 1 is that data ethics can be integrated into a local I/DMS and hence shouldn't be a problem integrating into the broader scientific community.

Pasta: A Network-level Architecture Design for Generating Synthetic Data Products in the LTER Network

- Mark Servilla, James Brunt, Inigo San Gil, and Duane Costa (LNO)

1 Introduction

The original call for LTER proposals in 1980 specifically addressed the need for comparative analysis of long-term ecological research at sites that represent major biotic regions of North America. Although comparative analysis is performed today, the process by which it happens is cumbersome and often ad-hoc. With the exception of ClimDB and HyrdoDB[1] [1,4], synthetic data sets are generated and managed by individual or small groups of investigators, along with their respective information managers, who provide some level of data quality and integrity. A consistent framework for synthesis across all 26 research sites is, however, non-existent.

In early 2005, Dr. Debra Peters of the Jornada LTER proposed the development of the "Trends" project [7] - "a large synthesis effort focused on improving the accessibility and use of long-term data." A major effort of the developing Trends project is to collect and integrate time-series data into a format that is consistent between the different data providers to facilitate synthesis. This process generally begins with a data request to individual researchers or information managers at the site. Data are generally returned as Excel spreadsheets or comma delimited text files, often without supporting metadata or information about their origin. This data is then transformed to a synthetic product through a number of processes, which may include converting site-specific units to standard units, performing quality checks and assurance, and mapping the site-specific format to a Trends standard format. Paramount to this process is the creation of metadata that describes every synthetic data product. Each step is time consuming and has the potential for error due to the repetitive human interaction during data manipulation.

Missing from the original project goals, however, is the ability to add into the synthesis process new time-series data collected from ongoing experiments. The LTER Information Management Executive Committee and Network Information System Advisory Committee recognized this deficiency and recommended the development of a prototype system to automate the synthesis process, and more specifically, to automate the integration of new time-series data. The following paper proposes both an architecture for generating synthetic data products within the LTER Network, as well as describes a prototype implementation as applied to the Trends project. This architecture is known by the moniker "Pasta"[2], and is being developed by the Network Information System group at the LTER Network Office.

2 Background

Each LTER site independently collects, documents, and archives their data for analysis and publication. In most cases, these data are made available for community scrutiny and further analysis after 2 years from the date of collection. Today, almost all data collected are documented by using the Ecological Metadata Language (EML), an XML-based data structure for describing scientific data that can be validated by using a community published XML schema. These EML documents are then uploaded into the LTER's data catalog, Metacat [2,5], a schema-independent XML database that is optimized for storing, searching, and retrieving EML documents[3]. The LTER Metacat contains just over 25,000 EML documents as of October 2006.

Although not required, LTER sites are recommended [8] to follow a strict document naming-convention for their EML that is comprised of a document *scope*, a unique *identifier*, and a *revision* number, joined together to form the string `SCOPE.ID.REVISION`. This name is referred to as the *Document ID* and provides a unique reference to the EML document within the Metacat database. The document scope is a string value that is the concatenation of the site's three letter acronym to the end of the broader domain identifier of "knb-lter", thus forming a complete scope hierarchy (e.g., `knb-lter-sev` is the scope associated with EML documents from the Sevilleta LTER, which is part of the LTER Network and part of the informatics research funded through the Knowledge Network for Biocomplexity (KNB)[4]). Both the identifier and the revision number are integer values. The identifier need only be unique to the site, while the revision number marks the version of the EML document being reviewed. A higher revision number is added to the Document ID if a change or modification occurs to either the data, the metadata, or both as documented in the EML. This number must be mathematically ordinal for all documents using the same scope and identifier.

To simplify and automate the upload process of EML documents to the Metacat, the Harvester application was developed as part of the KNB project. The Metacat Harvester allows individuals to register *http accessible* EML documents into a database table, along with the frequency of which the set of documents should be checked for updates. If the version of the EML document found at the site is more recent then the version found in the Metacat database, it is automatically copied from the site and inserted into the Metacat database. To date, the LTER Metacat Harvester application automatically checks for updates of more than 5,000 EML documents from 23 sites.

Both EML and Metacat are fundamental components of the Pasta architecture, providing a metadata standard capable of fully describing synthetic data products and a database system optimized for storing and retrieving such metadata.

3 Architecture

The Pasta architecture is based on a hybrid data warehouse model, which collects and organizes distributed data into an integrated data store. The following section illustrates the Pasta architecture through a dissection of its work-flow process, from the data source at the site to the end storefront where "synthetic" data products are made available to the community. The work-flow allows us to partition the architecture into individual modules that can be succinctly described performing their unique task. In fact, each module is designed to be independent of the others, and can be replaced with a new module that provides the same functionality without impacting the entire system.

For clarity, we define data that is extracted from the site as "source" data. The physical structure of how source data is represented at the site, as defined in the site EML document, is the "local" schema. The local schema is replicated in the Source Database. Data that is produced by transforming source data is defined as "synthetic" data. The physical structure used to store synthetic data is the "global" schema. The global schema makes up the framework of the Synthetic Database.

3.1 The Work-flow

The Pasta work-flow can be characterized by dividing the architecture (Figure 1) into six separate steps, each of which can operate independent of one another, but as a whole perform the goal of automated data synthesis. The process begins at the individual research site and ends at the storefront that exposes the synthetic data to the community.

The work-flow steps are:

1. Site data collection, quality checking, and EML metadata generation.
2. Harvesting new or revised site EML by the Metacat Harvester and inserting into the Metacat.
3. Recognizing new or revised EML documents that are specifically registered as part of the synthesis process. These EML documents are then parsed into their entity and attribute objects, which are then used to create a new database table in the Source Database. Finally, the data is extracted from the site and loaded into the new table.
4. The source data are then transformed from their local schema (captured by the Source Database table) and into the global schema as a new data product in the Synthetic Database.
5. The synthetic data product is documented with new or revised metadata, which is then inserted into the Metacat database as an EML document.
6. The synthetic data product is then made available for the consumer through a warehouse storefront, either a web-browser or web services interface, which accesses the Synthetic Database.

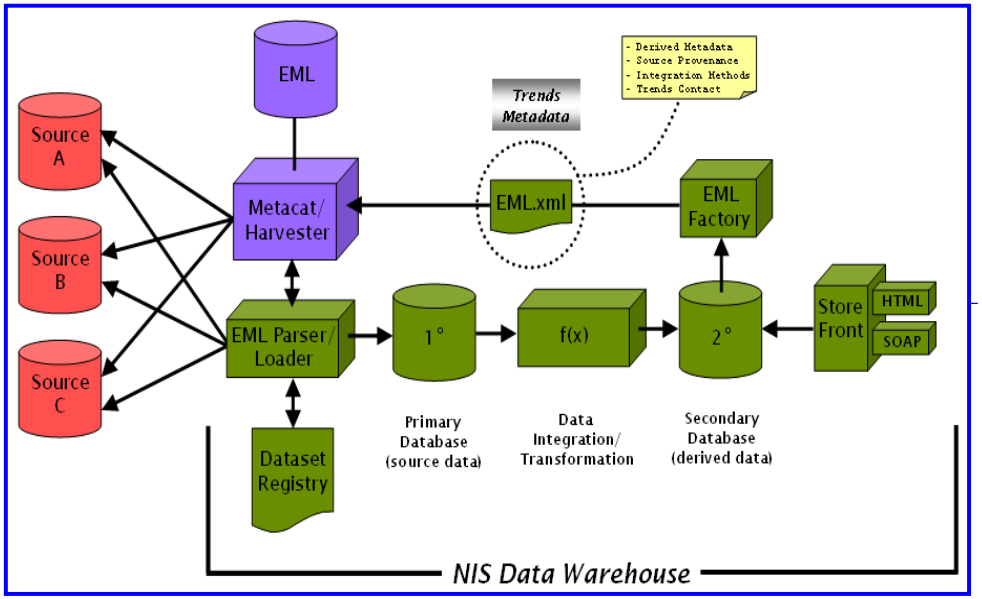


Figure 1: Major components of the Pasta architecture.

3.2 Steps 1-2: Site Data Management and EML Harvesting

Although steps 1 and 2 of the above work-flow are not technically part of the Pasta architecture, they are significant modules, and their absence would prohibit the execution of the complete work-flow. For this reason, they are described as pseudo-modules of the architecture.

As described earlier, LTER sites document all aspects of their data by expressing their metadata content in the Ecological Metadata Language. A revised version of the EML document is created if, for example, new data is collected and added to the package (as with time-series data), data was modified to correct for errors, or if any metadata content has changed (e.g., the collection end-date or contact information has been revised). Changes to the EML revision number in the Document ID will cause the Metacat Harvester to copy the EML document from the site and insert it into the Metacat database. Older versions of the same document are deprecated from the active system, but are never deleted completely. Such deprecated versions become an integral part of the overall package provenance. This process, represented in Figure 2, is ongoing within the LTER Network and operates independent of the Pasta architecture.

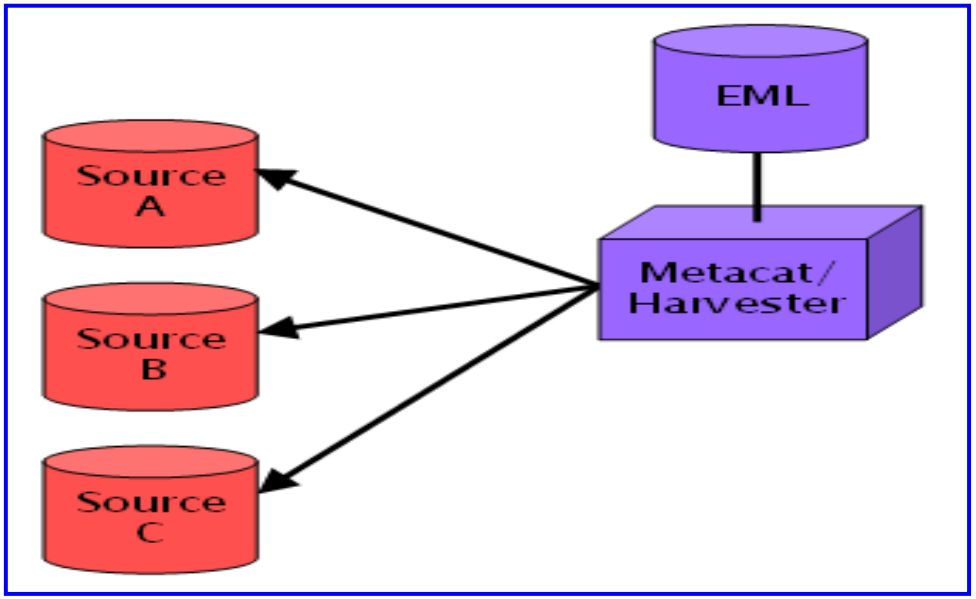


Figure 2: Work-flow steps 1 and 2 corresponding to architecture components for Site Data Management and EML Harvesting.

3.3 Step 3: Extraction and Loading

The third work-flow step includes the Dataset Registry, EML Parser, and Loader components (Figure 3) of the Pasta architecture. This module is responsible for identifying, parsing, and loading site data into the Source Database.

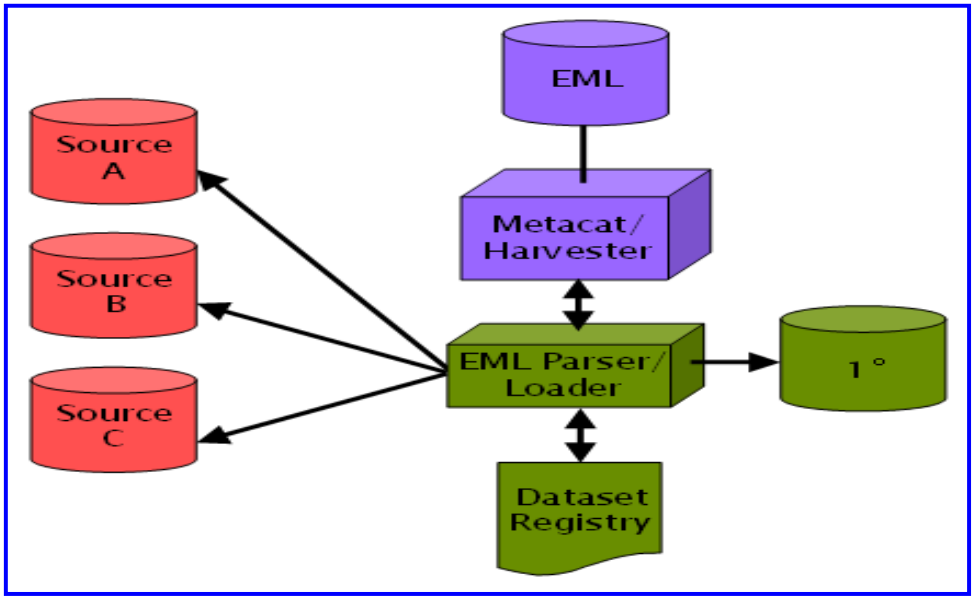


Figure 3: Work-flow step 3 corresponding to architecture components for identifying, parsing, and loading site data into the Source Database (1 [?]) .

3.3.1 Dataset Registry

The Dataset Registry is used to register source data that are used during the synthesis process. Technically, it is the EML Document ID that is recorded in the registry, along with information relating source data to synthetic data. Specifically, the registry identifies what source data are used to create a synthetic product and the appropriate routine that performs the transformation. The current implementation of the Dataset Registry uses PostgreSQL as its relational database.

3.3.2 EML Parser

The EML Parser is a Java-based library that reads and parses EML documents that are specified in the Dataset Registry. There are two critical functions of the EML Parser: 1) to create a new table in the Source Database representing the local schema defined in the EML document and 2) to open a connection to the site's data store and copy the referenced data into the new table. The development of this library is a collaboration between the LTER Network Office and the National Center for Ecological Analysis and Synthesis. The library, when complete, will become part of the Ecological Metadata Language software distribution.

3.3.3 Loader

The third component of this module, also a Java application, is the EML Loader. The Loader acts as the controller process between the Parser and Dataset Registry. Its primary responsibility is to read the Dataset Registry and compare the revision number of EML documents in the registry to those in Metacat. If the revision number in Metacat is greater than the one found in the registry, the Loader will assert a new extraction and loading sequence by calling the appropriate functions in the Parser. In this prototype, the Loader is also responsible for invoking specific transformation routines of the Transformation Engine and calling the EML Factory to generate a new EML document for the synthetic data. These two steps occur immediately after loading new data into the Source Database and are only necessary to continue the work-flow; they are not critical to the successful operation of this module. This module is still under extensive development and testing.

3.4 Step 4: Transformation

Transformation is the fourth step of the work-flow and defines the "Data Integration and Transformation" module (Figure 4) of the architecture. This module performs the mapping of data from the local schema, as represented in the Source Database, to the global schema through the use of a Transformation Engine - a collection of programs that perform the mapping operation (labeled as f(x) in Figure 4). We assume that individual transformation routines are *a priori* aware of both schemas, including input and output data types, and their semantic meaning, thereby eliminating the need for more advanced knowledge-based reasoning algorithms.

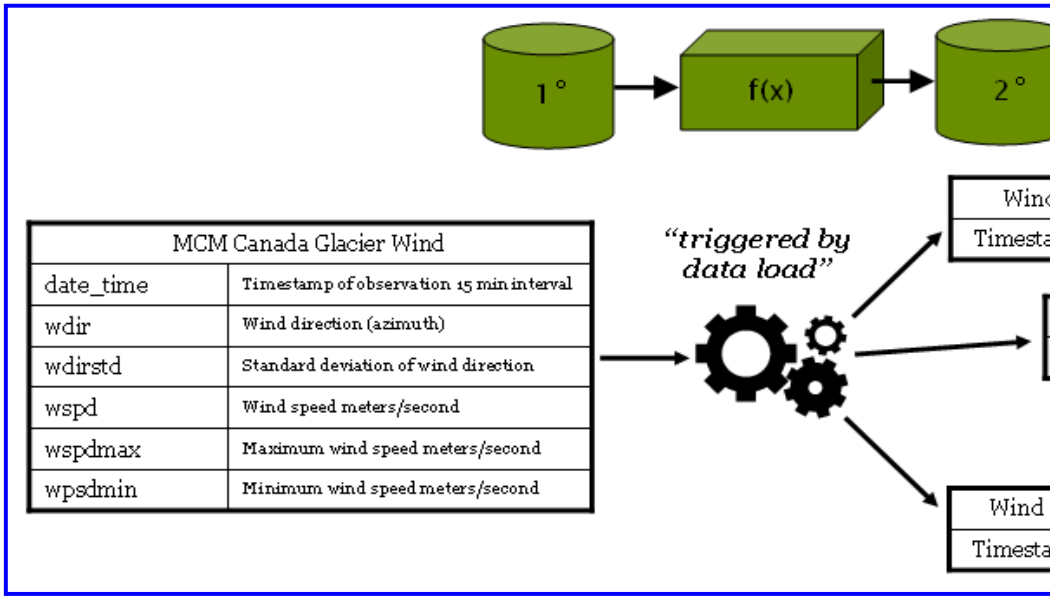


Figure 4: Work-flow step 4 captures the transformation of the local schema in the Source Database (1 ?) to the global schema in the Synthetic Database (2 ?) through use a Transformation Engine (f(x)).

3.4.1 The Global Schema and Synthetic Database

For the Trends prototype, we have constrained our working data model to point-location time-series data, such as climate observation data (e.g., a fixed sensor at a known geographic latitude and longitude that collects temperature values every 30 minutes). Such data is common within ecological observatories and provides a simple model to design and evaluate the architecture. As such, the global schema data table consists of only two important attributes - the observation *time stamp* and the *value* being recorded[5].

Mapping from the local schema to the global schema consists of three basic steps. First, tables that are composed of multiple attributes are separated into individual tables containing only the time stamp of the observation and the observation value of interest. This step is illustrated in Figure 4 where a local schema table describing weather data from the McMurdo LTER is separated into product specific tables of the global schema. Second, observation values are converted to the appropriate standard unit. This step may include data type conversion (e.g., integer to float) or an actual conversion from a non-standard unit (e.g., Fahrenheit) to a standard unit (e.g., Celsius). Third, the time stamp and observation value are scaled to comply with the time-scale defined for the synthetic data product (e.g., hourly to daily or monthly). Further processing, such as combining two or more source data into an aggregate product, may take place at this point.

It is important to note that execution of the Transformation Engine for synthesis of new or revised data results in a new data table being created in the Synthetic Database. For these tables, we follow the same document naming convention used for EML - that is, `SCOPE.ID.REVISION`. To manage the synthetic data tables, the global schema uses two additional support tables - the "product" table and "revision" table (Figure 5). The "product" table contains scope and identifier information for the synthetic products and the "revision" table records the lineage of product revisions. With these two support tables, any synthetic data product may be identified and accessed, including those that have since been deprecated by a new revision. For the Trends project, the table name scope is `knb-eco-trends"`[6].

Additional tables may be added to the global schema to support ancillary functions of the system, including to provide content for the EML metadata documents. The Trends prototype has tables to hold information about the station location (where data is being collected) and biotic relationships between individual synthetic products, neither of which are critical to the operation

of the system.

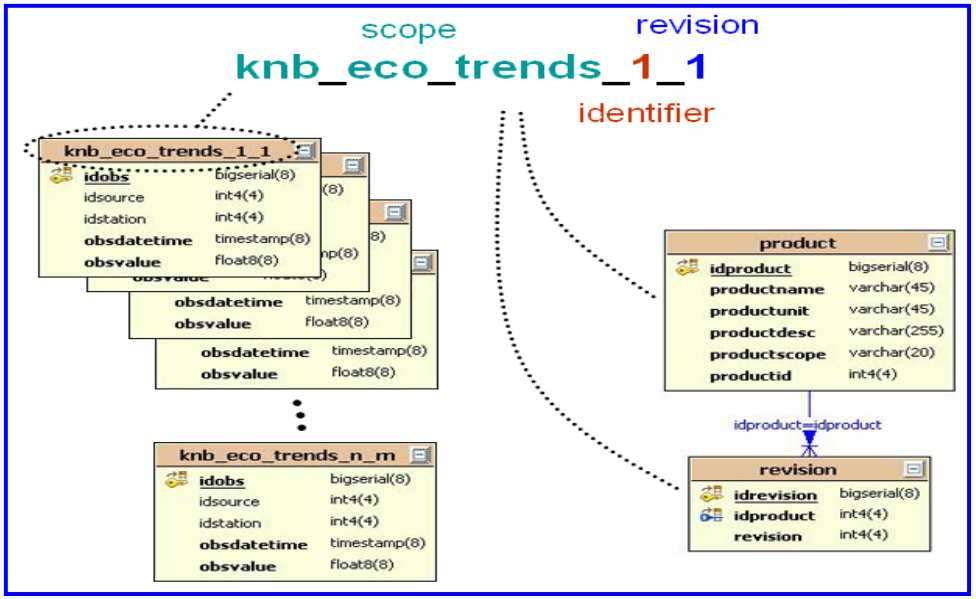


Figure 5: Global schema entity relationship between synthetic data product tables and support tables.

3.4.2 Transformation Engine

The Transformation Engine is simply the software code that performs the mapping between data in the local schema to data in the global schema (i.e., the synthesis process). It is noteworthy that transformation routines have the flexibility of being written in any programming language that support reading and writing relational database tables.[7] In fact, individual routines can be written in different languages. Our current implementation utilizes the "R" environment for statistical computing, and has only been tested in a few examples. In this case, individual R programs are invoked from standard shell scripts that are referenced from the Dataset Registry. The R language provides database support, including ODBC and SQL operations, both of which work well with the PostgreSQL RDBMS version that we use for the Source and Synthetic Database.

3.5 Step 5: EML Factory

The EML Factory module (Figure 6) of work-flow step 5 supports the metadata documentation process for all synthetic data products. Planned as a Java-based application, the following section describes how the EML Factory is envisioned to operate.

The EML Factory would be invoked as a side-effect of creating a new synthetic data product. The current design uses the EML Loader component as the calling process. An alternative is to have a database trigger invoke the EML Factory module, but this capability is not available in all relational database applications. Yet, another possibility is to have each transformation routine execute the EML Factory directly after completing its processing.

Content for metadata is divided among four physical entities: 1) plain text files storing static content that changes infrequently, such as project management information or the LTER Data Access Agreement; 2) database tables of the global schema storing dynamic content that is directly associated with the synthetic data product, including date ranges, data statistics, and online access information; 3) programming code that is used to create the synthetic data product; and 4) the EML metadata of the source data product.

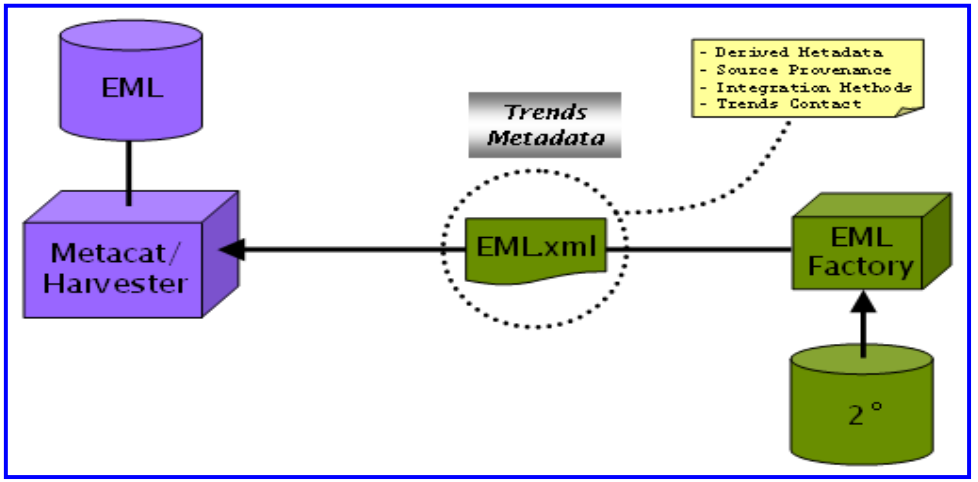


Figure 6: EML Factory module of work-flow step 5.

The EML Factory would use a predefined XML schema template to generate the EML document fields. It would then use metadata content from the four entities described above to complete the EML document. Static metadata from the plain text files and the dynamic metadata from the database would be placed in the canonical fields defined for such data in the EML specification. These fields include those found in the "resource" group and the "dataSet/dataTable" subtree of EML. Both the programming code for the transformation routine and metadata content from the source data EML document would be stored in fields of the "methods" subtree (Figure 7). Specifically, a "methodStep" element would be used to capture both types of metadata. In the case of the source data's EML, minimal content must be copied directly from the original document to the required fields of the "resource" group within the "methodStep/dataSource" element, but the remainder (which, in most cases, is the complex metadata) can simply be referenced by placing the appropriate Metacat URL for the EML document in the "online" element of the same "resource" group. The final EML document would then be inserted into the Metacat for access through either the Storefront or the standard Metacat interfaces.

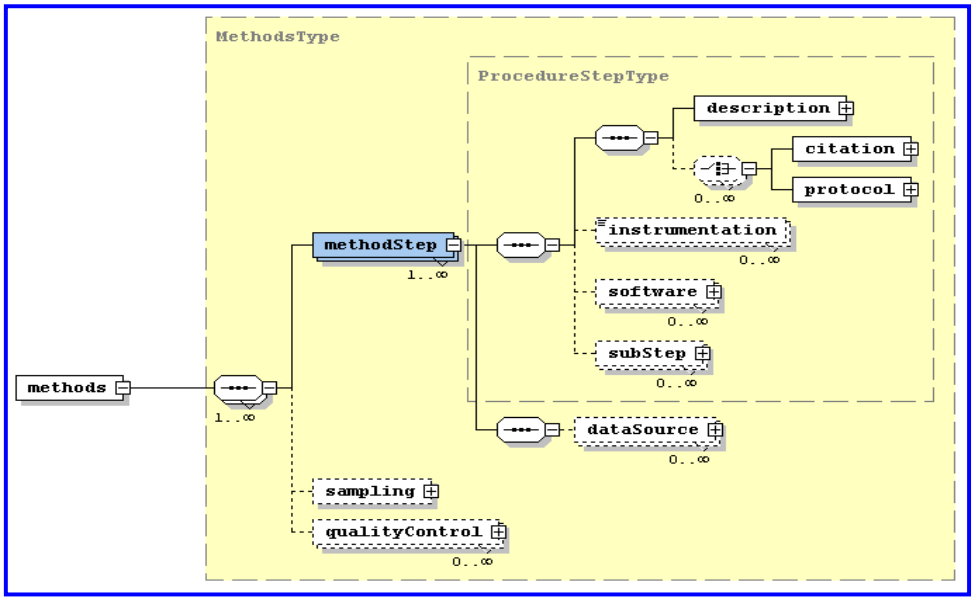


Figure 7: The "methods" subtree of EML 2.0.1 as viewed in XMLSpy 2005.

3.6 Step 6: Storefront

The final step of the work-flow is the "Storefront". The Storefront module of the architecture defines the community access point to the Synthetic Database and its content. The core table structure of the Synthetic Database, including the product, revision, and data tables, forms the primary scope of interaction to the synthetic data. Ancillary tables may be accessed to develop a well rounded human-interface that provides a variety of query tools to interact with both the data and the metadata stored in the Synthetic Database and in the Metacat as EML.

The current interface of the Trends prototype consists of a website that connects to the Synthetic Database for data access and the Metacat for displaying the corresponding EML document in its native XML format. The website provides a basic query interface that allows the user to select one or more of the synthetic data products for viewing. Once selected, the website displays a brief summary of metadata, along with a dynamically generated plot of the data (Figure 8). The metadata includes links that lead to site information, a data download page to directly access the synthetic data, and a product lineage list that displays all revisions of the selected product.

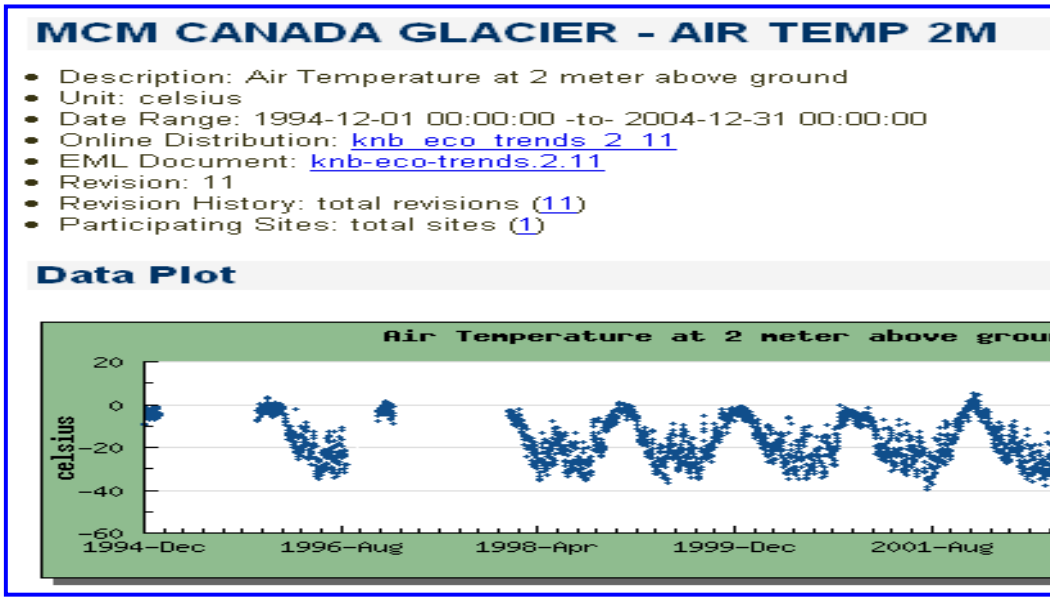


Figure 8: Example screen shot of work-flow step 6, a synthetic data product with plot.

The website utilizes the PHP server-side scripting module for the Apache web server to provide programmable logic (i.e., database connectivity and program execution control) to the overall web site. The plotting application, which is an add-on PHP library called "JpGraph", is tightly-coupled to the web server. We also explored the use of more loosely-coupled packages that reside outside of the web server scope - specifically, we tested the efficacy of calling an "R" script from PHP to generate the plots. Although this approach seemed to work reasonably well, due to time constraints, we opted for using the JpGraph library for the prototype.

Our plans include the development of a web-service interface over Simple Object Access Protocol for direct access to synthetic data and metadata. This aspect of the project is still under early development.

4 Discussion

Although the Pasta architecture is still in the design and prototype phase, its current blue-print provides a solid foundation for meeting the goal of generating Network-level synthetic data products. Its overall design incorporates a number of

salient features, including 1) the use of data warehousing strategies for data integration and archiving, 2) providing support for tracking data lineages and provenance, 3) a design that easily integrates new replacement modules with little or no impact on other components of the system, and 4) a simple and compact global schema that leverages already proven informatics tools used within the LTER Network and broader community.

4.1 A Data Warehouse Approach

Data warehouses are focused around a centralized store of information that conforms to a physical global schema [6,10]. Data from outside of this store must be mapped from their local schema to the global schema through a transformation process. It is common within a data warehouse to use the phrase "extraction, transformation, and loading", or by its acronym ETL, to describe the process in which data external to the central store is "extracted" from the source location, "transformed" to the global schema, and "loaded" into the final store database[8].

Another common approach to the integration of external data is to provide a virtual interface, or view, that "federates" each site's data store to appear as if a single store is being accessed [3,9]. Data warehouses differ from federated systems by enforcing the mapping of external data to the global schema by loading and transforming the data before being accessed by the consumer. In this case, the transformed data reside in a store that is within the domain of the warehouse. A federated system, in contrast, relies on dynamic mapping of the external data to a more loosely-coupled "mediated" schema through a common software interface, such as a web application.

Issues of data quality, including *persistence* and *freshness*, often arise between the two approaches. Data warehouses tend to promote persistence of data by acting as an archive, while federated systems are characterized by providing the most up-to-date data available to the consumer. From our perspective, the data warehouse model is more aligned with a science-based synthesis environment where reproducibility and open access to all data are critical. Our vision here is that any site data loaded into the Source Database will be permanently archived (although, data that is never or rarely used would be moved to near-line or off-line backup media), thus making it always available to the synthesis process, and ultimately the consumer. A federated system cannot meet this level of service since it operates outside of a site's functional realm and cannot guarantee data availability (e.g., one or more sites may be unavailable). The North Inlet LTER site, no longer participating in the LTER Network program, is one example of a site whose data is no longer available for synthesis[9].

The most up-to-date data, on the other hand, is best served by a federated system. The Pasta architecture achieves reasonable "freshness" by utilizing the revision concept of the EML to proxy a site-based release schedule. In this instance, the latency of freshness may be quantified by summing the time between EML revisions at the site and the frequency of the Harvester schedule.

4.2 Data Lineage and Provenance

In this context, we define data *lineage* as the linear progression of a specific data product over time, as in a sequence of revisions that are generated by adding new time-series data to a product line. It is assumed that each revision is preserved as a discrete "snap-shot" in time, thereby having a fixed start and end date. Similar to lineage, we define data *provenance* as the origin or history of a specific revision. In other words, provenance records the series of steps that generate a data product, such as the transformation process that produces synthetic data, including the origin of the source data. We can illustrate these two concepts as a matrix (Figure 9) that shows lineage as a progression in time from t=0 to t=n and provenance as a product changes through transformations from S to S^m, with D as the final synthetic data product. We call this the "Heritage Matrix". Both lineage and provenance are crucial for scientific data integrity and reproducibility. Lineage provides the ability to work with data from a previous state in time, while provenance allows that data to be reproduced from its origin.

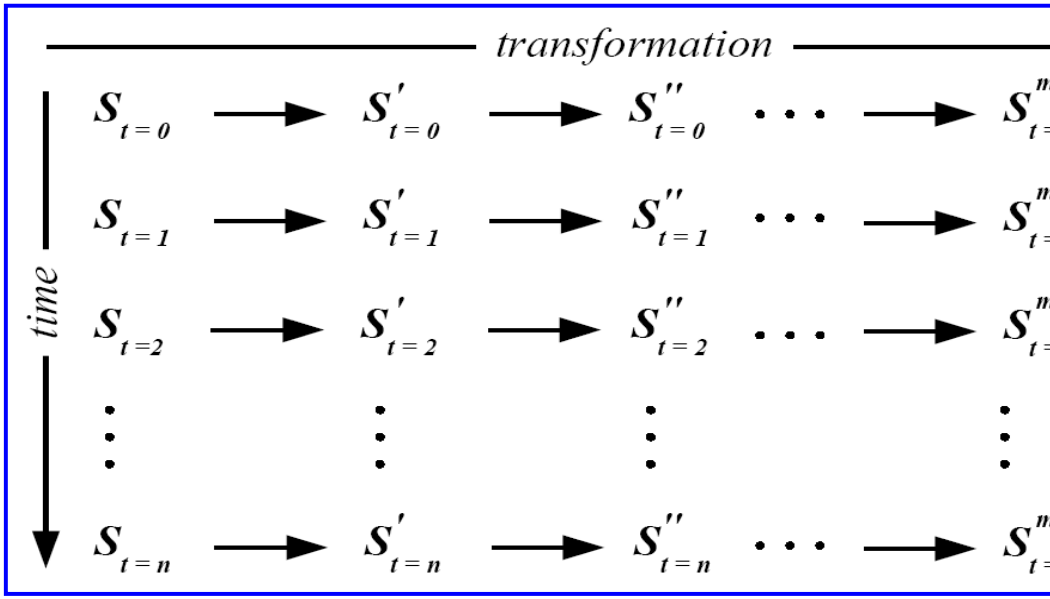


Figure 9: The Heritage Matrix showing both lineage and provenance as progression in time and product change, respectively.

The Pasta architecture supports both lineage and provenance. Lineage is simply the by-product of keeping each data table that is created for every new revision of a synthetic data product. The lineage list can be seen in the Trends prototype when viewing "product details". Since revisions are never deleted, a user will always be able to access an earlier data product. This is especially important if overlapping data values change from one revision to the next.

Provenance information is recorded in the EML document for the synthetic data product. There are two parts of the provenance record as described earlier: 1) the software code of the transformation routine that is used to produce the synthetic data and 2) the content of the source data EML document, including the reference URL of the specific document located in the Metacat.

4.3 Module In, Module Out

The overall design of the Pasta architecture is defined by separate and autonomous modules that can operate independent of one another (as opposed to a single monolithic system that tightly-couples each process). This provides a great advantage in the development and testing of the modules since individual modules can be built in parallel. Front-end test cases can be easily developed for specific modules so that they may be demonstrated for functionality and user acceptance without relying on those modules that are part of the early work-flow steps. An example from the Trends prototype is the development of the Storefront module, which was designed, implemented, and demonstrated using a manually generated testbed Synthetic Database for data access. Because the global schema will remain consistent between work-flow steps, we presume with confidence that the Storefront will operate continuously and unknowingly when the full system is put into place.

This modular approach also has the benefit of simplifying new module integration into the future. As new techniques for data integration and synthesis become available, individual modules can be replaced without concern for the other components of the system.

The opposite is also true. Individual modules can be taken out of the Pasta architecture and used else where in the ecoinformatics community. An example of this is the EML Parser library, which will become integral to both the Pasta architecture and to Metacat when it is complete.

4.4 Extensible Global Schema

The global schema of the Pasta architecture is unique due to its compact nature - the core schema requires only a single data table structure (although, many actual data tables) and two support tables. This compact nature is possible because much of the content that most systems would require to store separately is already embedded in the EML documents. As a result, the global schema is quite portable from one database management system to another, and allows more superficial schemas to be constructed around it, thereby extending the overall functionality of the Synthetic Database. The current Trends prototype uses PostgreSQL and provides minimal content beyond the core global schema. We plan to move this instance of the global schema to the MySQL database system and augment the content with additional metadata for the completed Trends project, thus enabling a more comprehensive query environment.

5 Conclusion

The Pasta architecture described above is part of an evolving suite of technology being investigated to address the challenges of Network-level synthesis. This architecture is only in its early stage of development, but promises to meet some long-standing goals of the ecoinformatics community - namely, an automated process to access site data and produce a synthetic data product that is available to the community.

The design and prototype development of the Pasta architecture provides only the foundation for automating synthesis at the LTER Network. Already, there are visions to expand the current architecture to take advantage of the work being performed by community-based information technology, like the informatics research taking place at the National Center for Ecological Analysis and Synthesis, and the super computing centers, including the National Center for Supercomupting Applications and the San Diego Supercomputer Center.

6 Acknowledgments

The Network Information System group at the LTER Network Office would like to thank the LTER Network Information System Advisory Committee, and especially Wade Sheldon and Don Henshaw, for motivation and direction on this project. We sincerely appreciate our collaboration with the National Center for Ecological Analysis and Synthesis, including the insight provided from Mathew Jones, Mark Schildhauer, and Jing Tao. A sincere appreciation goes to the Trends project team and the editorial committee, with special thanks to Debra Peters, Christine Laney, and Ken Ramsey. Work completed on the Pasta architecture is funded by the National Science Foundation under Cooperative Agreement #DEB-0236154.

References

[1]Baker, K., B. Benson, D. Henshaw, D. Blodgett, J.H. Porter, and S.G. Stafford. 2000. "Evolution of a Multisite Network Information System: The LTER Information Management Paradigm", Bioscience, vol. 50, no. 11, pp. 963-978.

[2]Berkley, C., M. Jones, J. Bojilova, and D. Higgins. 2001. "Metacat: A Schema-Independent XML Database System", 13th Intl. Conf. on Scientific and Statistical Database Management, p. 171.

[3]Halevy, A.Y. 2001. "Answering queries using views: A survey", The VLDB Journal, pp. 270-294.

[4]Henshaw, D.L., M. Stubbs, B. Benson, K. Baker, D. Blodgett, and J.H. Porter. "Climate Database Project: A Strategy for Improving Information Access Across Research Sites", presented 9 August 1997, workshop on Data and Information Management in the Ecological Sciences: A Resource Guide, Albuquerque, NM.

[5]Jones, M.B., C. Berkley, J. Bojilova, M. Schildhauer. 2001. "Managing scientific metadata", IEEE Internet Computing, vol. 5, no. 5, pp. 59-68.

[6]Lenzerini, M. 2002. "Data Integration: A Theoretical Perspective", PODS 2002, pp. 243-246.

[7]Peters, D. and C. Laney. 2006. "Trends in Long-Term Ecological Research Project", Jornada Trails, Jornada Basin Long-Term Ecological Research Program, vol. 10, no. 1, p. 2.

[8]Sheldon, W., J. Brunt, D. Costa, C. Gries, J. McGann M. O'Brien, K. Ramsey, and M. Servilla. 2004. "EML Best Practices for LTER Sites", working document, LTER Network Office, 21 p.

[9]Ullman. J.D. 1997. "Information Integration Using Logical Views", ICDT 1997, pp. 19-40.

[10]Ziegler, P. and K.R. Dittrich. 2004. "Three Decades of Data Integration - All Problems Solved?", WCC 2004, pp. 3-12.

Footnotes:

1Both ClimbDB and HydroDB are Network-wide synthetic data products for site climate and hydrological data.

2The name "Pasta" originates from a passing comment made regarding the original project title, "Phase II". The comment, "*Phase II* is about as informative as *Spaghetti and Linguini*", resulted in a more tangible noun, "pasta", but without any more meaning. It has since been accepted as our internal architecture moniker.

3Refer to http://www.ecoinformatics.org for more information on the EML specification and Metacat.

4Funded by the National Science Foundation under Grant No. DEB99-80154.

5There are, of course, additional attributes allocated for data table management.

6This scope string is only a place-holder for testing purposes.

7In the current prototype, however, they must also be callable by the EML Loader component.

8The Pasta architecture actually performs extraction, loading, transformation, and the final loading.

9The North Inlet LTER data is currently archived at the LTER Network Office.

Three Challenges in Supporting Shared Workspaces

- Mason Kortz (CCE)

In the past months the Ocean Informatics team at Scripps Institute of Oceanography has worked to create shared workspaces for the PAL and CCE sites. For a discussion of the technologies behind these shared spaces, and their place in a collaborative infrastructure, see the Spring 2006 Databits issue (Kortz, 2006). The initial setup for these shared

workspaces required design work. Below are some of the challenges we faced, along with our thoughts on how they could be solved, mitigated, or at least prepared for. The ideas below are not necessarily best practices, or even strong recommendations - they are rather local lessons learned about many challenges, and potential solutions, that exist in supporting shared workspaces.

1. Defining the purpose and scope of a shared space. In a collaborative environment, especially one with multiple shared workspaces, it is important to define the role of each shared space - its purpose and its scope. The purpose of a shared space may be archiving, dissemination, collaboration, or some combination of purposes. You should also consider whether policy or technology will limit use of your space to one type of content, such as raw data, documents, or images. The scope of a shared space is typically defined in terms of its user base. The user base may be a group of individuals, a pre-established organization, or simply 'everyone'.

Once the role of a shared space has been defined, that information must be made available to the user base. Providing a purpose for a space prevents the 'what-goes-where?' syndrome that often leads to shared spaces becoming disorganized, misused, or abandoned. Providing a scope helps users avoid posting sensitive material in a public space, or putting information in a place where the intended recipient cannot access it.

The process of defining shared spaces can also be illuminating for administrators. If two or more spaces are defined as having the same purpose and scope, then it may make sense to merge them into a single shared workspace. Alternatively, you may find that a particular purpose-scope combination is not present, which may prompt further questions about the use cases of shared workspaces in your environment.

2. Insuring integrity of shared information. In a shared workspace, many users may be working within the same space and editing the same information. Frequently, in shared environments, especially in those designed for collaborative work, users will have privileges that allow them to alter or remove the information created by other users. Because of this, there is an increased need to protect the integrity of shared information. An advanced shared space could support version control - the ability to track the addition, deletion, and editing of information, as well as to revert to a previous state if an unwanted change is made. In some shared spaces, the underlying technology offers this support; in other cases, an ad hoc change-tracking system can be implemented as a set of best practices.

Shared spaces that support version control at a technology level provide a robust solution for information integrity. Almost any change collision can be sorted out with no loss of information. However, such systems can be restrictive, in that the technology used defines what type of shared space is created. For example, website content management software can provide a version-controlled space, but only for the purpose of creating web pages. Further, version control technology creates additional overhead, as it essentially adds a new layer into any interaction with your shared space. Version control at the technology level makes the most sense for spaces with a focus on collaboration, where frequent change collisions are possible.

In some cases, implementing version control at the technology level seems to be over-engineering. Both in a space where collaboration is not the focus, and in a collaborative space with a small enough scope, a small set of best practices can be defined. Best practices that encourage users to avoid altering others' work and to leave a record of changes made essentially work as an ad hoc versioning system.

3. Maintaining consistent support for multiple users. A shared space, which by definition must support many users, can raise issues of consistency. One of the most obvious, but also potentially most frustrating, is the issue of supporting consistent access to your shared space. Because this depends on both the technology that supports your shared space and the platforms the user base uses to connect to the access the shared space, there are a great number of variables to take into account. Testing access to a shared space before launching can catch many of these problems, but changes to the user base or your infrastructure are always possible.

Other issues with consistency arise from practices, rather than technology. With many people accessing the same

space, different choices will be made in terms of organization and nomenclature. Enforcing organizational structure can be difficult, and with smaller shared spaces it often isn't necessary. When it is necessary, it is best to anticipate the needs of the space and create the structure ahead of time, rather than relying on users to create it as they go. Similarly, there are many cases in which enforcing nomenclature is unnecessary, as long as information is readable to the user browsing it. In cases where information may need to be sorted or searched by a program, however, a strict nomenclature should be used.

Example: WebDAV Share. Many of the issues above were encountered while setting up a WebDAV file sharing space for both the Palmer and California Current Ecosystem sites. We began by considering the purpose of the shared space. We needed a space where the entire research community could work jointly on various projects. As such we defined the purpose as quick, easy collaboration and file exchange. Because of technological limitations with WebDAV, we did not have the option for a granular permission system, so we added the caveat that this space would not be used for sensitive material. We expanded this purpose to include use of a general dropbox, by which researchers could get data, publications, and other information to other researchers and the information management team. We defined the scope as all participants associated with a site (i.e. researchers, graduate students, technicians, outreach, and administrative staff).

Next, we considered what we could do to ensure data integrity. The WebDAV server was already part of our backup system, so disaster recovery was covered. For short-term data integrity, we considered software version control through the application Subversion, but decided that the additional layer of complexity did not fit the 'quick and easy' formulation of the shared space's purpose. Instead, we decided to implement a few best practices. First, when editing another user's information, do not overwrite their work - create your own copy of the resource and edit that. Secondly, when making additions or edits, include your name and a timestamp (using a specified format) in the file name, so that other users can quickly identify the most recent version of a shared resource. Thirdly, don't delete shared files; instead, move them to the 'to be deleted' space, which is regularly cleared of files older than one month.

With these ideas in place, we implemented the shared space. Early on we noticed a cross-platform issue: Macintosh OS X users could upload files, but not folders. We realized that the server was blocking the hidden .DS_Store files in the Mac folder, causing the copy operation to fail. We also added some structure to the shared space, creating four directories at the top level: data, metadata, individual, and other. We created further structure in the data and metadata directories so that researchers would know where to upload files, and created a convention for uploaded data file names. The other directories we left as an open area without any enforced structure or nomenclature. Finally, we created a README file in the top level of the share to detail the purpose and scope of the shared space as we had defined it, as well as the best practices and conventions we had created. The file also lists the names and email addresses of the users that with access to the share. This README file also contains instructions on connecting to the shared space, and is emailed out to new users when they are granted access.

The WebDAV spaces, as well other shared workspaces, are now part of the PAL and CCE information infrastructure that supports collaborative work. As we both expand and refine our support for collaborative work, new challenges arise that are both technical and social in nature. As with many shared tools, technical foresight and explicit best practices - both gained from our experiences in the past - help to ease the introduction of these shared resources into scientific practice.

Ocean Informatics Matlab Working Group, Mirroring the LTER Community Approach

- Lynn Yarmey (CCE/PAL)

Matlab is a matrix manipulation software application that meets the computation and graphics needs of many data handlers in a number of different research groups. A new Matlab Working Group (MWG) represents a mechanism for sharing experiences, building community and shortening the learning curve associated with this powerful software package. The group meets a growing need for local technical training and communication within an academic research organization. MWG is developing within the Ocean Informatics environment, which is formed around the long-term arrangements of the CCE and

PAL LTER sites in collaboration with the California Cooperative Fisheries Investigations (CalCOFI) and other Integrative Oceanography Division participants at the Scripps Institution of Oceanography.

The group was founded in June of this year after recognizing a recurring thread in technical data analysis discussions: there are difficulties encountered by new users starting out in Matlab. New instrumentation and data streams are overwhelming traditional local data handling capacities involving legacy 2D, Excel-based processes. As a first step toward preparing for future workflow applications, Matlab is a reasonably priced and robust alternative, but the transition to Matlab with its software environment, structured line commands, and matrix concepts is daunting without training, resources or help. Previous local Matlab communities waned with changes in infrastructure, needs and personnel. With a new influx of individuals starting out, this community working group provides a mechanism to link up with other new or more experienced Matlab users in the physical and organizational vicinity.

The current working group consists of more than a dozen active participants as well as a number of inactive members. Membership crosses social boundaries by including staff, students and researchers. Participants are involved in a wide range of interdisciplinary projects and all come with a diversity of datasets and interests. Once a month, with a different member preparing and presenting a particular topic of interest, we meet informally to discuss, share experiences, and mentor those not familiar with the material. Meetings are theme-based, with previous topics including methods of importing data into Matlab, debugging scripts and functions, and an overview on the concepts and uses of cell arrays. Rotating the meeting presenter ensures relevance, promotes engagement, prompts leadership and creates a mechanism for participants to delve deeper into targeted topics.

By building interdisciplinary partnerships amongst the participants, the group presents a unique opportunity to design and enact local variable naming conventions, coding guidelines and common workflows, which all hold the potential of contributing to a shared infrastructure. Organizationally, we are in the process of designing a community script repository to pool home-grown tools and resources, building upon the group's enthusiasm for better documentation and linking personal coding methods. Through team dialogue about concepts, problems and solutions, we are developing a common language that can help us articulate our needs, work and accomplishments as data collectors and users, which in turn inform researchers in their data handling plans and data scoping practices. Also, the group creates an arena for building identity, prompting for feedback, and redesigning communication processes. For example, we initially discovered that Matlab resources (books, blog posts, etc) were scattered and not easily available. In response, we have developed a group website that brings many of these resources together along with locally pertinent information and presentation products from our meetings.

The group is moving slowly and in a personal and grounded way towards goals including generic data ingestion, pre-processing, processing and initial visualizations. This process of loose standardization involves community decisions based on an intimate knowledge of local data acquisition, handling, practices and common needs (i.e. specific oceanographic calculations, corrections, etc). In many ways, technology has created an obstacle to this process rather than simply adding capabilities. Matlab bugs, multiple operating system representations and a range of computing and display capacities have all served as limiting factors in bringing together the data and data processing of individuals. We are working to mediate these differences through development of platform-independent coding guidelines and optimization training as well as by making local solutions available in the form of a shared script repository.

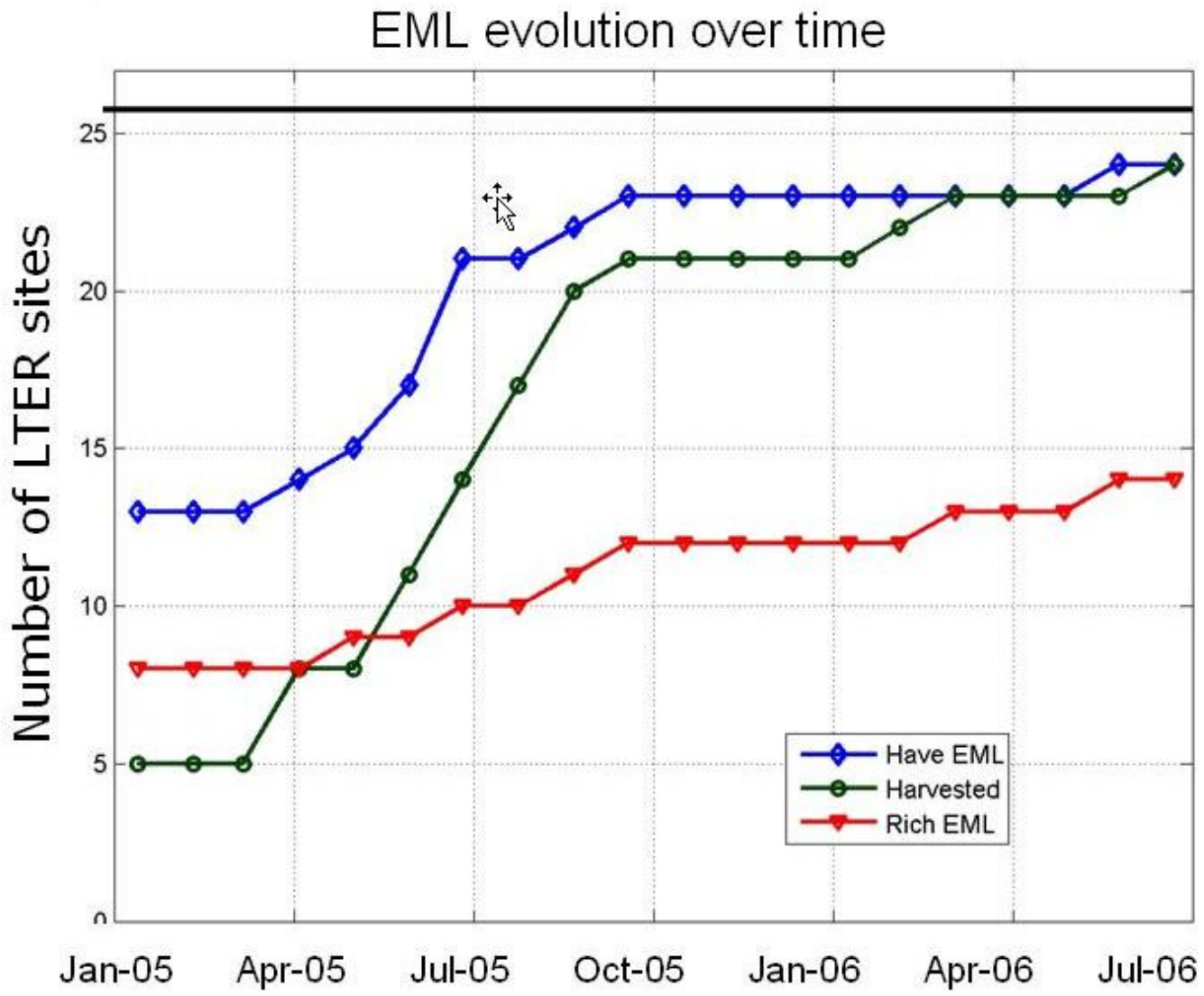
The Ocean Informatics endeavor focuses on creating an environment supportive of collaboration and mutual learning. The Matlab Working Group contributes to such efforts as a community approach to exchanging, preserving, building and sharing knowledge. WIth an emphasis on tools, partnerships and ongoing education, MWG mirrors a number of the features of the LTER community that support site science.

EML Status of the LTER Sites, Data Synthesis Examples and the Next Step

- Inigo San Gil (LNO)

LTER Information Managers recently attended the Information Management (IM) meeting in Estes Park, CO as part of the 2006 LTER All Scientists Meeting, where we enjoyed a week of interaction with many information managers from the LTER network, as well as members of the International LTER IM community. We also benefited from the close contact with LTER scientists, many of whom are supported by the work of IMs.

In a short twenty minutes, I had the opportunity to outline the status of where LTER sites are in terms of the metadata standardization project. In summary, all of the LTER sites, with the exception of the two new sites (MCR and CCE) and Palmer, have implemented the EML as an standardized form of metadata publication to the community. Furthermore, about half of the LTER sites are offering very rich EML documents through the LTER Metacat and its sister servers. The same 5,000+ metadata documents are also available in the Biological Data Profile standard through the very fast network of servers at the National Biological Information Infrastructure. This remarkable achievement places the cadre of LTER sites among the leaders of the larger scientific community in terms of data and metadata management.



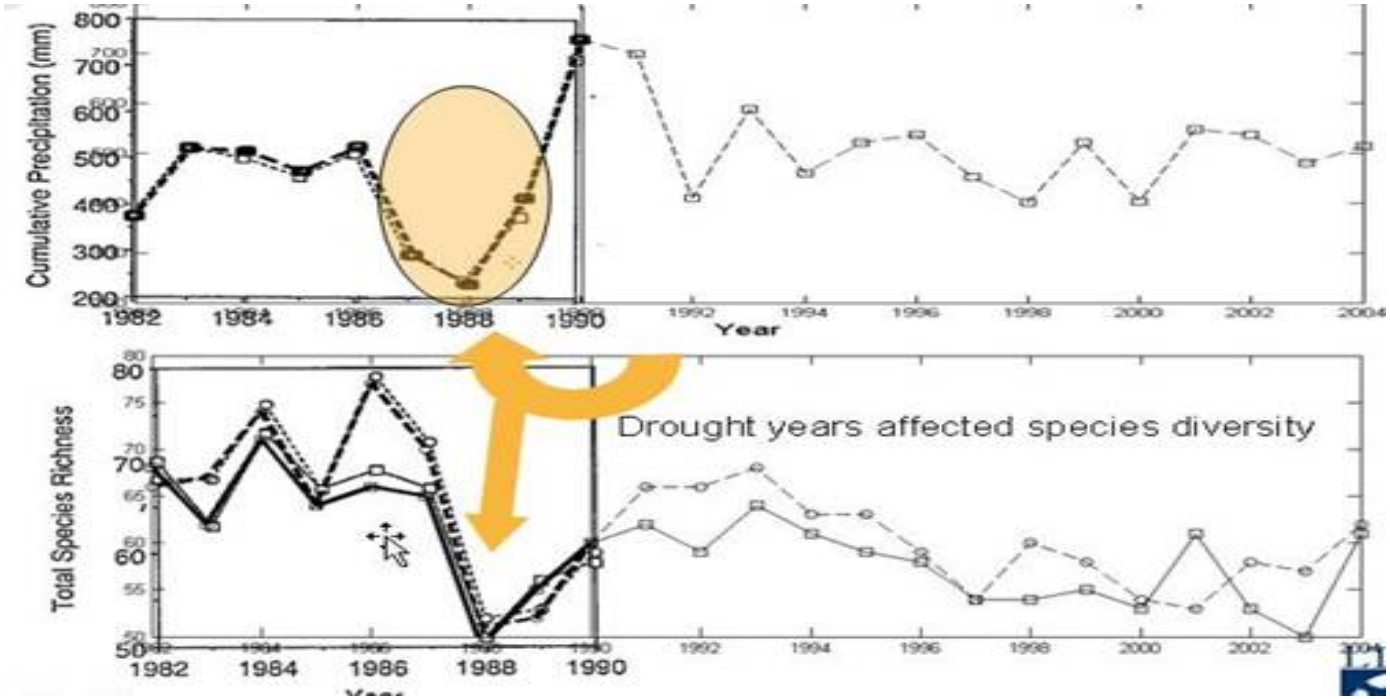
Time evolution of the EML standard adoption at the LTER network. The number of sites adopting EML increases over time, until reaching the limit of 26, the total number of LTER sites. The red curve shows the sites that offer rich EML, while the other two curves show the sites that implemented EML (blue curve) and offered EML in Metacat (green curve)

Examples of synthesis of LTER data

Perhaps the most exciting part of this brief report was to demonstrate a handful of application prototypes that performed data synthesis by exploiting rich EML documents. One example of such a prototype is the web store of the TRENDS project. This dynamic data graphing web store relies heavily on EML for data digestion and manipulation, as well as data and metadata provenance.

Other example of synthesis tools that are based on rich EML is the NCEAS synthesis tool Kepler. Kepler is a scientific tool to create workflows for data synthesis, which is in the second beta release at the time of writing.

We also demonstrated a simple and highly customized prototype to replicate published results based on McMurdo Dry Valley's meteorological data *(Climate Cooling Revisited1)*. Results from a similar prototype tool were showed at the 2006 Organization of Fish And Wildlife Information Managers meeting (OFWIM). In this case, Stephanie Lyon (CDR) and San Gil (LNO) used biodiversity data gathered at Cedar Creek Natural History Area *(Biodiversity in Grasslands1)*. We should not forget that these are prototypes -- a work in progress. These examples, however, demonstrate that rich standardized metadata is not just a pretty utopian package lacking pragmatism; it is a requisite for powerful data synthesis tools.



This graph replicates and expands time evolution of biodiversity data published by Tilman et al (Oecologia, v89 257-64, 1992). The results of the paper were replicated exactly and the expanded graph corroborates the published results. A study on the Antarctic Climate Cooling was shown at an ASM poster, reproducing and expanding the original results in a similar fashion

Next steps in EML implementation

What is our next step? There is the mammoth task of performing systematic quality control tests on our existing EML. It is not enough to reach attribute level EML (level 5 EML or rich EML). We need to make sure the information expressed in EML is accurate -- in other words, that it really describes the associated data! The EML standard itself can be improved in many ways as exemplified by the number of Bugzilla entries (there exist many EML issues!). At the same time, we should do our part in making sure we provide the best metadata we can get. To this end, the LNO NIS team (Servilla, Costa, Brunt and San Gil) and others are working on processes to ensure quality assurance of data and metadata. Such tools will facilitate the QA/QC labor of each IM. Note that a related task is the QA/QC of the data itself. A post-ASM workshop has been proposed to address data quality checks by Sheldon (GCE), Henshaw (AND) and Ramsey (JRN).

I would like to thank all the IM community for their hard work, as well as the LNO staff for their dedication to these tasks. I encourage you to continue the effort and lead the broader scientific community and data managers towards high quality metadata standard.

Special thanks to Duane Costa and Mark Servilla for their valuable help in editing and organizing this article.

Ecological Metadata Language-Based Statistical Analysis

-John Porter and Chi-Wen Hsiao (VCR)

As diverse as ecological data types are, the users are even more diverse in the approaches they take to the analysis of data. Some users depend heavily on spreadsheets, some use GUI-based statistical software such as Statistica and JMP while yet others use text-programmable statistical packages such as MATLAB, R, SAS and SPSS. We also expect to see the growing use of Kepler and other scientific workflow systems in the future.

Regardless of the system used there are some basic steps that are followed when conducting an analysis (Table 1). At each step there are some roles that can be filled by EML documents, and tools needed to achieve the desired results.

Table 1 Analysis Step	Use of EML	Potential Tools
Ingestion - raw data (often as a text file) is read into specialized, internal forms used by the statistical software, such as a spreadsheet, a SAS Dataset, an R dataframe or an SPSS system file.	The dataTable and physical modules of EML can be used to automate this process, including the handling of missing values (missingValueCode), and production of labels (attributeLabel) for variables/ attributes and for coded values (codeDefinition).	The PTAH project has produced basic translators for EML to SAS, R and SPSS. (see the Spring 2006 Databits). The Taiwan IMTeam has produced tools for web-based analyses using "R" and for displaying locations using Google Map and Google Earth.

Quality Assurance (QA) - data needs to be checked to make sure it is read in properly.	EML can provide information on ranges for numerical variables (bounds). Additionally summary outputs giving the frequencies (categorical variables) or means and ranges (interval and ratio variables) can be produced.	The PTAH tools incorporate code to produce QA reports.
Integration - in this optional step, data from several different sources may need to be transformed and combined.	EML unit module can provide some insights into which variables/ attributes in different data sets might be candidates for combining. Similarly, unit can help to determine the appropriate transformation (e.g. Fahrenheit to Celsius) for standard units. However, ultimately the researcher must decide what is appropriate.	The SEEK and KEPLER (http://www.kepler-project.org/) projects are pursuing tools for automating integration of data. Basic web-based tools could also be developed.
Analysis - depends on the specific questions to be addressed by the research.	The potential use of EML for automation is very limited, since the desired analyses are a function of the research objectives rather than a function of the dataset itself. However, the extensive textual components of EML describing methods and research context can be invaluable.	Currently the display pages provided by METACAT are the best source of information on methods. However, more specific display tools aimed at putting the needed documentation at the fingertips of the researcher are needed.

The usefulness of EML for automation is primarily in the first two steps (ingestion and QA), where a "rich" EML document can provide all the needed information. More sophisticated tools using EML as input can semi-automate the data integration process, but displaying candidate attributes for combining, along with suggested transformations.

The degree of support afforded to researchers by different tools varies. There are also limits imposed by the content and quality of EML documents and the statistical software that make some of the applications less "robust" than is desirable for researcher use. In specific:

- EML does not contain the information needed to actually access the underlying data. The EML physical module does contain online elements such as url that could be used to point to the data. However, since most LTER sites require information on who is using the data, the link is usually to a web form - not the data itself. We need to develop a web-services infrastructure that will support authentication, so that EML can truly provide access to the underlying data.
- EML documents vary in the quantity of information provided and on its quality. At the very least, an EML document needs to include a dataTable along with an attributeList and a physical element to be able to support automated data ingestion. For QA, additional information on missing values and bounds are required. Additionally, this information must reflect what is actually in the underlying data file. During testing we have encountered many inconsistencies, from extra quotation marks built into EML documents, to attributeLists whose order did not correspond to the actual order in the file, to incorrect delimiters being listed. These problems would apply not only to automated analysis tools, but also to researchers manually creating statistical programs. Given the complexity of EML and the large number of datasets to which it is applied such "growing pains" are not unexpected. In fact, attempting to use the automated analysis tools are a great way to do QA on the EML documents themselves.

- Some statistical programs and databases are not very robust with respect to "surprises" in the data or are intolerant of certain conventions. For example, in data prepared for use in analysis by SAS the preferred form of a "missing value code" for a numerical variable is a string, typically a single character (e.g., 'M'). However, the same data when processed with "R" or most database programs will generate a fatal error, since a character string is not allowed in a numerical field. Inconsistent use of headings in delimited files can cause similar problems. We need to develop some "data filters" that can identify, and rectify these sorts of problems before the data is passed on to analytical programs. One approach might be to do automated ingestion into a database (where all the rules are enforced) and then have the database generate files in custom forms suitable for meeting the conventional requirements different analytical software.

None of the problems listed above are insurmountable, and indeed overcoming them is a normal part of the evolution of EML-based ecological information systems. However, overcoming them is a must. To be used by most researchers, ecological analysis tools need to be robust and failure tolerant.

Incorporating Information Management into the Luquillo LTER Schoolyard Program

- Eda C. Melendez-Colom, Elliot Lopez, Jorge R. Ortiz-Zayas, Clara I. Abad, Hilca Nieves Rí¯³ and Aurea Berrí¯³ Sá¥º (LUQ)

This information was presented as a poster at the LTER 2006 All Scientist Meeting (http://luq.lternet.edu/outreach/schoolyard/Activities/ASM2006Poster_IncorporatingIMLUQLTERSchoolyard2.ppt)

Scope. The Luquillo Experimental Forest Long-Term Ecological Research (LUQ LTER) site has a long history of outreach activities that includes developing educational web sites on the subject of ecology. In particular, the Luquillo-LTER Schoolyard Program in Puerto Rico (LUQ LTER Schoolyard) is the product of a joined effort from several governmental agencies and federal-sponsored institutions (Table 1).

Table 1. Participating institution and roles in the LUQ-LTER Schoolyard Program

Agency or Institution	**Active membership and roles**	
Department of Education, Puerto Rico	Teachers	Educate students and coordinate them in the field trips to collect data
	Teachers' coordinator	Coordinates outreach activities and communicates with teachers
Forest Service International Institute of Tropical Forestry (IITF)	Field Technician	Trains students and teachers in field methodologies
	Director	Provides scientific background in field methodologies and analysis.
University of Puerto Rico Institute for Tropical Ecosystem Studies (ITES)	Science coordinator	Provide administrative and scientific guidelines
	Outreach coordinator	Directs the group in their outreach activities
	Information manager	Assist in the entering of data and metadata, developed and maintain web site, provide computer trainings to schools

In this poster we focus on the components and activities of the LUQ LTER Schoolyard program emphasizing the recent inclusion of an Information Management component. This has added the technical support and structure that the schools' communities needed and has further encouraged them to communicate their technical and administrative issues. In addition, closer interaction between the LTER scientists, the Program Information Manager with the teachers and students has enhanced the quality of data collection and manipulation.

Introduction.
The LUQ LTER Schoolyard in Puerto Rico shares the goals with the Department of Education of Puerto Rico. Our efforts are directed to educate youth to develop a curiosity to study nature and to enunciate environmental problems, and their possible solutions. This is made possible by field experiences that provide the student with technical knowledge and learning processes for developing research skills in data collection and analysis techniques.

Schools.
Three public high schools in Puerto Rico are participating actively in the Luquillo LTER Schoolyard Program, some since 1987. Each of them has its own research goal while conducting similar studies with the same standardized methods to gather and manage the data. Table 2 gives a description of each school's main research sites and project goals.

Table X1. Schools sites' description and main project's goals

School	Study site description	Project principal goal
Pablo Colon Berdecia High School in Barranquitas	• Montaí¥†na Torrecilla located in Barranquitas, PR • 942.75 m (3,093 ft) asl • secondary forest • a coffee plantation from the 1800's to 1940 • Manata Rí ’iver's origin • study area divided in 2 50X50 m2 plots with 25 10m x 10m quadrants	How fast our trees growth after Hurricane Georges?
School	**Study site description**	**Project principal goal**
Juan Ponce de Leon High School in Florida PR	• Mogote Cuba in the municipality of Florida, PR • kart zone • secondary forest • Single plot of 2000 m2	Study the habitat of uncommon species such as *Coccoloba pubescens*
Francisco Morales High School, Naranjito	• Anones Ward in Naranjito, PR at the borderline with the school • secondary forest • coffee plantation abandoned in the 1950's, today a partial plantain cultivation • two research plots of 2500 m2 each	Develop diverse projects on environmental health and conservation education

Organizational Methods.
The LUQ LTER Schoolyard follows the LTER model of integrating people from different sites in working committees with

the variant that its active members are from different institutions, not all from the LTER program. Table 1 contains a list of these institutions and the roles of the members of this committee.

Standardized Methods and Information Management. Two main sets of standards are used to promote cross site comparison and enhance the quality of the research:

- standardization of the methodology used in the field;
- standards used for managing the information.

Field Methodology.
Assisted by the International Institute of Tropical Forestry (IITF), the three schools established their study areas (see Table 2) and started their data collection applying a similar methodology. Table 3 describe some relevant methods; Popper et. al., (1999) gives a full description and a summary table that has been published in the LUQ LTER Schoolyard Web site for the schools' rapid access and public reference (http://luq.lternet.edu/outreach/schoolyard/default.html).

Table 3. Field methods used at the Barranquitas and Naranjito schools as described in Lugo (1999)

Measurement	Methodology
Trees' with =4cm dbh and understory < 4cm dbh	Diameter measured at 1.3m height
Trees height (=13.5m)	Height measured with calibrated pole
Taller trees	Range finder

Information Management was done differently by the three schools, unlike the field methodology which was coordinated and monitored by a single agency that sets standards. Although all schools carefully documented their methods and measurements, this was done in three different ways using different formats of documentation (word documents and documentation integrated in Excel spreadsheets) thus representing a difficulty in the comparison and access of these information. Several steps have been taken toward standardization, presented in Table 4.

Table 4. Standardizations steps performed toward the schools data management procedures (.../ = (http://luq.lternet.edu/outreach/schoolyard/)

Procedure or document	Reference or explanation
Protocol documents prepared and published in the web site	(../Documents/MeasurementProtocols-allSchools.htm)
Naming conventions assumed	These includes: common species names, one sites' similar data variables names, file naming
Similar data structures when doing the same kind of measurement	Same order and labels given to data sets with same kind of measurements; inclusion of a date column
Data entry sheet example	Sheet for entering the plant measurements developed as an example for other kind of data
Same set of metadata standards	Following the LUQ LTER metadata guidelines

These standardizations will facilitate future cross-site comparisons and studies.

The documents are downloadable from the Program's web site: http://luq.lternet.edu/outreach/schoolyard/default.html.

Activities and Plans.
A series of activities have been made and planned to train students and teachers and to inform the community of the program's achievements and goals. Table 5 lists these activities.

Table 5. List of activities of the LUQ LTER Schoolyard

Planned or past activity	Goal or achievement
"Understanding the Arcabuco" - Internship in Nov 2006	Students will experience through the complete process of research from formulating a science question to data gathering and manipulation, and finally analyzing and reporting data..
Teaching Ecosystem Complexity through Field Science Inquiry Workshop - 2006	Teachers reported on the elements of the program that have been most beneficial to them to learn about ecology, field research, how to use this knowledge and these skills with students.
Information Management Workshop - 2000	Teachers where introduced to the concepts of LTER Information Management and where presented with entering and analyzing tools
"Cumbre de Bosque" - Symposium 2000	Students and teachers were exposed to the studies of Puertorriican scientists and students from the three schools presented their LTER Schoolyard studies.
Water Information for a Sustainable Environment (WISE) Project - 2000	Guidelines were presented to less privileged communities on the best use and conservation of water and sewers ("sumideros")connecting the underground watersheds underneath Florida PR. Students were awarded by the Thames Waters Co. for their initiative.
First Environmental Workshop for Parents - 2000	Parents where presented with the LTER Schoolyard students research studies and new technical resources. Workshops on water pH determination and atmospheric studies were offered to parents.

The LUQ LTER Schoolyard committee identified the need for more technical support and better ways of communication. Information Management is now being used as a resource to directly train students and teachers in the concepts and techniques used by the LTER sites. A teachers' workshop on LTER Data Management and an online static directory was completed in 2000. Table 6 presents a summary of these activities:

Table 6. LUQ LTER Schoolyard Information Management and support activities

Teachers and students training in data management concepts and best practices

Design and maintenance of the LUQ LTER Schoolyard Web site (http://luq.lternet.edu/outreach/schoolyard/index.html)
Definition of a local mail server's alias or mailing group (lterschool@ites.upr.edu)
Metadata standards development and completeness
Data entry and quality control
Evaluation, enhancement and standardization of data sets structure of all schools
Hands on workshop on the entry, quality control and data preparation for analysis in an internship that will be held at the research station on Nov 2006

Summary.
The LUQ LTER Schoolyard program represents an interagency effort that integrates different disciplines. The presence of an information management component has stimulated and facilitated the professional interaction among the members of the program and will facilitate future cross-site activities and synthesis efforts.

Acknowledgements.
The authors are grateful to the support provided by staff from the University of Puerto Rico Institute for Tropical Ecosystem Studies, particularly to Eva Cortezand Carmí® Quijano. The Puerto Rico Department of Education, through its Science Program has supported this initiative and has facilitated the participation of teachers and students. Private land owners have allowed access to the study plots showing a remarkable commitment to education and conservation. Mr. Carlos Dominguez and the staff of the USFS International Institute of Tropical Forestry has been instrumental in the development of the field program and has maintain a high level of motivation among teachers and students. Dr. Jess Zimmerman and Dr. Steven McGee have also provided insightful direction and assistance. . Finally, the authors thank Dr. Ariel Lugo, who's vision has allowed the development of this program for more than two decades.

References.
Popper, N., C. Domingí®§uez Cristió¢¡¬, A. Santos, N. Melendez Irizarry, E. Torres Morales, A.E. Lugo, Z.Z. Rivera Lugo, B. Soto Toledo, M. Santiago Irizarry, I. Loucil Rivera, L.A. Zayas and C. Coló®® 1999. A comparison of two secondary forests in the coffee zone of central Puerto Rico. Acta Cientí¦©ca 13:27-41.

LUQ LTER Schoolyard Web site: http://luq.lternet.edu/outreach/schoolyard/index.html

Editorials

A Long-term Investment of an LTER Information Managerâ€™s Time

- Eda C. Melendez-Colom (LUQ)

In these past seventeen and a half years that I have been working as an information manager, I have seen many of my colleagues leaving. This is not surprising considering the opportunities to switch jobs out there. So the inevitable question is: why have I stayed this long? It is certainly not for the pay and neither it is because the job is an easy one to do. How

easy could it be when I always have the feeling of being behind and that I still have a lot to learn and accomplish in order to catch up? It does not matter how ahead of the game I think I am, there is always something or someone that reminds me how much more I have to learn. This feeling is so overwhelming sometimes that I think I should simplify my life and stay home baking, a task for which I also have a special talent.

Well...one part of it is the challenge, the sense that I will never feel that I know it all, that I have to have the disposition to learn more, because there is always more to learn. This forces me to stay energetic and interested. This makes me feel young!

Then there is this certainty that I am doing something important; that in every activity I perform I am contributing for the future of science, of humanity. It is my vision that the major contribution of the LTER Program to humanity is the mentality it forces upon the people that do the science. Sharing and doing collaborative work is the biggest and most important seed that, to my understanding, the LTER has planted in different communities. It is the present and the future.

We, the LTER information managers are very aware of this principle. We facilitate that sharing and by doing so, we engage in numerous activities that make us ambassadors of the LTER Program. Also by doing so, we face both the resistance as well as the willingness of the members of the different communities we deal with to do that sharing.

I have always sensed the openness of our LTER students in absorbing the new concepts and ideas of the LTER program when I interact with them. Unfortunately, sometimes they are buried in their own survival problems and / or have a very direct influence of a non-willing scientist or professor that cannot see the profit or self-advantage of sharing, not even their metadata, leave alone their data.

Once in a while there is this student who has become a scientist and surprises me with the display of that bloomed seed I did not even know I planted. This student is so aware of the importance of documenting and sharing data that they feel they have to share this knowledge with the rest of their peers.

Last week this happened to me at the Nov 2006 LUQ Monthly Meeting. Late in the 1990's, I had offered a workshop on data management and the use of Paradox (a RDBMS) to enter data to the students of one of LUQ's long-term projects. A student who has now finished her PhD work and is doing a post-doc at the International Institute of Tropical Forestry (IITF) attended this workshop. While managing her thesis data she recalled the data management concepts she learned back then and expressed to me her willingness to organize a workshop on information management for the present LUQ LTER graduate students. The idea is to make them aware of the many difficulties they would encounter when dealing with complex data bases if their databases are not well designed and planned for.

She attended the 2-weeks workshop held the first two weeks of October at La Selva field station in Costa Rica by the Organization of Biological Field Stations (OBFS). At our meeting this month, she presented, along with her thesis work, a synopsis of the presentation on metadata that was given to her at the workshop (by John Porter). In this presentation she talked about the different tools she was exposed to and gave a summary of their utility and use. Her presentation on metadata brought up a discussion among the scientists present at the meeting on why scientists would want to share their metadata and data. With all the years of experience giving presentations to scientist on this subject, I could have not accomplished what she did in those 20 to 30 minutes of presentation and discussion. When scientists give that much attention to the subject, you know that they are at least giving some consideration to the matter discussed.

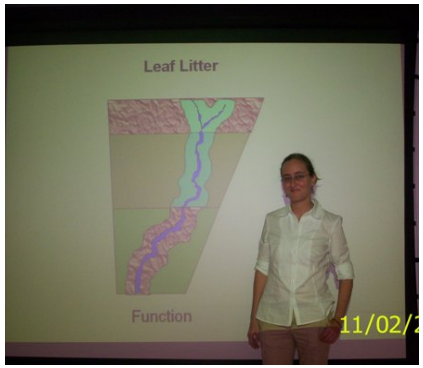


Figure 1: Tamara Heartsill-Scalley presenting her thesis data at the Nov 2006 LUQ LTER Monthly meeting

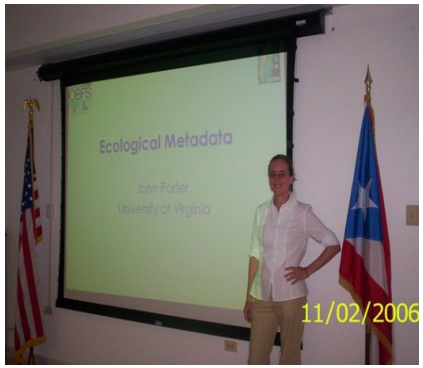


Figure 2: Tamara Heartsill-Scalley presenting metadata concepts at the Nov 2006 LUQ LTER Monthly meeting.

It was not only the satisfaction of knowing that I have been able to spread the LTER seed but the realization that we are making a long-term investment when we agree to teach our students. I have always been willing to invest time in students. They are the best users in filling out LUQ's metadata forms; they understand my guidelines with little or no intervention from my part; I guess this is because they are willing to invest their time in doing so. After this experience I am certain that I am really investing my time to help creating future LTER scientists.

Scientific Meetings: Rigor, Relevance, and Variety

- Karen Baker (PAL/CCE)

Scientific publications come in a variety of forms with different levels of rigor in peer and editorial review. During 2002, LTER Information managers participated in the Systemics, Cybernetics and Informatics (SCI) conference, in association with the LTER Information Managers Meeting. At two sessions organized by the LTER Information Managers, 12 papers were presented and also published in the voluminous conference proceedings (http://intranet.lternet.edu/committees/information_management/sci_2002/). This conference was selected because it was large (1,000+ attendees), international and multidisciplinary in character and provided a venue for peer-reviewed publication at a time when few journals dedicated to environmental and ecological informatics existed. However, questions about the scientific legitimacy of SCI have been raised when "nonsense" papers (in one case written by a computer program) were submitted and accepted (http://news.bbc.co.uk/2/hi/americas/4449651.stm; http://www.timesonline.co.uk/article/0,,11069-1571285,00.html). This has led some to propose removing all references to SCI from LTER web pages and publication lists, while others suggest this would be an overreaction, because inspection of the published papers in SCI indicates that most have scientific value. Here are addressed some broader issues regarding meeting venues and a suggested response to this controversy.

Meeting Dimensions

Meetings are a community resource with multiple dimensions such as organizational structure, information exchange, community building, and themed networking as well as scientific scholarship, publication, and domain validation. SCI addresses the first four of these in providing for conference structure, session dialogue, engagement opportunities, and community organized sessions. The latter three elements are considered in more detail below.

1. Scientific Scholarship and peer-review: A review process provides quality control for manuscripts. Paper acceptance at the SCI conference can take one of two forms, reviewed or non-reviewed. In the SCI organizational scheme, session chairs are key to manuscript integrity in terms of carrying out the review process. With the paper acceptance policy now provided

online, one can argue that the SCI review process is flawed from the outset in 1) accepting papers based on an author's CV and 2) providing a non-reviewed paper acceptance especially if it is not publicly distinguished from reviewed papers. The SCI procedure is to accept papers as non-reviewed when there is no review process conducted. Thus, when session chairs who do not carry out their responsibilities coincide with participants intent on fraud, there is no scientific validation.

Because the LTER community handled the review process of papers in our sessions, the scientific scholarship of our collection of publications stands on its own. Not only did our active participation create a combination learning opportunity and training event for information managers who had not previously engaged in publication, review, or international conference presentation activities, but also we were able to ensure the thoroughness of review in carrying out the process according to professional conventions.

2. Validation and sanctioning: In considering validation, it is important to recognize there are many sources of validation, ie individual, institutional, organizational, professional society, community familiarity or even 'success' itself. The LTER Information Manager (IM) meeting is associated with the LTER, an ecological community. When a meeting is held independently, it is no less valuable but for the IMs is validated as an LTER meeting rather than via association with another organization. When an LTER IM meeting is held in conjunction with an associated meeting, some crossover exchange and potential collaboration occurs. It is valuable to ask as we do periodically at our annual meetings: 1) What are associations of interest? and 2) What are the venues that stimulate and facilitate learning? An ecologist could respond with interest in associating with the Ecological Society of America. A computer scientist might point to Very Large Databases (VLDBs). A digital librarian might say the Digital Library for Earth System Education (DLESE). An organizational theorist might say Digital Government Organization (DGO). An information scientist might mention the American Society for Information Science and Technology (ASIST). They are all good responses. For an information manager, there is no single overarching source of validation but rather diverse associations and sources of validation.

3. Publication and journal ranking: Ranking of publications in and out of academia is an ongoing, contested, and valuable exercise debated and carried out separately by each individual, field, and academic unit. In addition, publication venues are currently in transition. For instance, unlike the past, new journals including e-journals are frequently not associated with commercial publishers. One way of classifying publications is by designating them as reviewed or non-reviewed. The proliferation of refereed journals brings forth a multi-tiered ranking ranging from top (A) journals such as Science and Nature to B & C journals In an academic arena where the review process insures high quality work, electronic and interdisciplinary publications are creating new situations - and bringing classification challenges along with new modes of information exchange. For instance, there are a burgeoning number of conferences and conference proceedings. Traditionally non-reviewed, these conferences are now frequently timely mechanisms of dissemination that are both reviewed and associated with professional organizations.

Question Responses

A set of summative LTER papers captured many aspects of the state of LTER information management in 2002. These papers continue to be available for other data managers, scientists and communities. The value of having prepared for and participated in the LTER sessions contributed to and remains a valuable part of our individual and collective experience bases. That is, the process was as valuable as the product. A few brief responses to other questions that arise:

- Will the LTER IM community attend a SCI meeting as a community again? Unlikely because the IM community has changed as well as the meeting and the publication landscapes.
- Does the tempest exposing the weakness of the SCI review system invalidate our publication collection or our experience? No.
- Should we regret having been part of this process? No.
- Would having a paper published one year in this forum have negative consequences on advancement of an information manager? Doubtful if there is not a multi-year trend. The full body of an individual's work is typically considered.
- Did having a paper published in this forum have a positive consequence in terms of conceptual advancement of LTER information management? Yes.

In Summary

The SCI 2002 forum provided a useful impetus to the LTER IM community work in terms of articulating LTER

information management issues, contributing to expanding a shared vocabulary, and developing local theory grounded in practice. Yet the SCI format seems to lack the capacity to converge a robust body of knowledge within a professional or commercial association; that is, it's very openness and inclusion brings with it a negative. Although the LTER IMs today have alternative forums and have associates of more direct interest to their work, one wonders about the value of loose, open forums for other emergent communities and professions outside traditional academic and professional support structures. Having a range of meeting types is valuable in any field to ensure a variety of forums are available to meet the wide range of community needs.

Given the issues involved, one method of acknowledging our work is to keep the public posting of the LTER paper collection but add a statement that clarifies circumstances. A draft of such a statement might read:

The following collection of papers represents work of the LTER information management community. The papers were presented in two sessions at the SCI 2002 meeting. Since the review process at this large scientific meeting has been demonstrated to vary, it is important to document that the review of these contributions was handled by LTER session chairs in accordance with traditional professional conventions including a minimum of two written reviews for each paper and mandatory revisions when necessary.

This is a time of transition in terms of dealing with recorded information. As members of the LTER information manager community-of-practice, we have some flexibility - even somewhat of a mandate - to explore new approaches and types of venues for information exchange and professional growth. Scientific rigor and scientific relevance involve a constant re-balancing. Choices are rarely black or white. The LTER IM community periodically chooses to reach outside-the-box technically, socially, and organizationally in order to adapt to changing contexts, capabilities, and understandings.

What influenced the balance and choices of meetings in 2002? It was less the need for validation than the need for broadening the environment within which to self-organize. Though an emergent profession with changing publication venues, we in the interim have things to say, experiences to synthesize, and stories to tell. A variety of forums and formats are needed. In taking the opportunity to look back to ask 'Is SCI a valid meeting?', let's also look forward to consider 'what are the needs of the LTER IM community?' and 'what types of meeting and publication arrangements meet the LTER IM community needs?'

IM Friendship and JaLTER

- Akiko Ogawa, Japanese LTER (JaLTER)

On the evening of September 18th, 2006 my name was announced as the next speaker in a small crowded room at the Annual LTER Information Managersâ€™ meeting in Estes Park, Colorado . I had just arrived after a long trip from Japan and hadnâ€™t fully switched to English-speaking mode yet. However, I felt less nervous than I thought I would because I could see several familiar faces in the audience. These were the faces of people, whom I made friends with not so long ago - Meei-ru Jeng, Tanya Yang, Yun Yin Yeh from Taiwan, Syuqiyah Abdul Hamid from Malaysia, Atzimba LÃ³pez from Mexico, Avinash Chuntharpursat from South Africa, and Peter McCartney, John Porter, Barbara Benson, and Don Henshaw from the US. I met these colleagues at two information management workshops held in Beijing and Taiwan in 2005 and 2006.

The East Asia-Pacific Regional ILTER Information Management Workshops were held as capacity building efforts for the members of the East Asia-Pacific Regional ILTER Network. At these two workshops, participants were introduced to various information management tools and methods by experienced IMs from the US, Taiwan, and China . This information would soon be used in framing the establishment of the JaLTER (Japanese LTER) information management system.

Currently, the information management systems at JaLTER sites are largely undeveloped. Even though Japan has a long history of ecological research and has an accumulation of long-term data for a wide variety of subjects, the data are not

typically published online. Many Japanese scientists have not yet embraced the importance of "metadata." Japan has lagged behind some other countries in implementing an information management system for two fundamental reasons. First, the idea of data-sharing has historically been very uncommon in Japan. In my personal opinion, which may not necessarily be shared with other Japanese ecologists, one possible explanation for the lack of data-sharing is that studies in Japan seem to focus on individual organisms. In contrast, ecosystem ecology, which has flourished in the U.S. and elsewhere, integrates many aspects of ecology thereby requiring data from a number of different sources. Secondly, there are no incentives that motive scientists to put an effort into information management, largely because information management in Japan is not yet required as part of research funding criteria.

To motivate Japanese scientists and funding agency officials, information management in Japan has to show its worth by strongly demonstrating its benefit to science. Just showing the "future" benefit of data management (e.g., data archiving) doesn't attract most Japanese scientists and funding typically only supports their "current" research and not "future" research. Therefore, it is critical to demonstrate more direct and immediate benefits of data-sharing and sound information management practices.

While JaLTER has a long road ahead before establishing an organized network IM system, having an undeveloped system has some advantages too. Scientists are more likely to agree on a standardized method since they have not developed individual preferences. For example, EML and Metacat may be accepted rather easily if they do not already have a metadata database. To facilitate the introduction of standard IM methods, tools must be made easy for Japanese IMs to use. Information management documents should be translated into Japanese and workshops should be provided to train people in the use of IM tools. One such effort I hope to undertake with a colleague, involves translating the book, "Ecological Data: Design, Management and Processing" by William Michener and James Brunt into Japanese.

JaLTER is on its way to becoming an official ILTER network member in 2007. The preparation committee is in the process of selecting candidate sites and drafting the network by-laws. The draft by-laws will include statements on the establishment of a network information management committee consisting of IMs from each site, and a requirement for all sites to have concrete plans for the development of a site information management system. I am hoping that the new JaLTER kick-off as an official ILTER member will promote and bolster both the JaLTER site and network information management systems.

"Data management is people management" and building relationships among IMs is an important part of establishing an information management system. During the IM meeting in Estes Park, I was impressed with the strong connection among site IMs of the USLTER network. I could see that this strong connection is the key to the large progress of the USLTER network information management system. I now share that connection, and many of the people that were strangers to me on the first night of the meeting are now friends of mine. These friendships were mostly made through informal conversations during lunch, Salsa dancing, and hikes in the woods. I believe that the various activities during the LTER-ASM and IM meeting will extend the collaborative relationships and friendships among USLTER information managers to ILTER network information managers. And such relationships will become a great help for JaLTER to catch up with other networks in the development of an information management system.

I appreciate being invited by the LTER Network to the 2006 IM meeting in Estes Park. I want to express special thanks to Nicole Kaplan and Kristin Vanderbilt who made all the arrangements to accommodate participants from overseas. I would also like to thank the other USLTER IMs who welcomed us with warm hospitality and friendship.

News Bits

Hubbard Brook Dedicates Archive Building to Cindy Veen

- John Campbell (HBR)

In July, the Cheif of the USDA Forest Service gave approval to name the Hubbard Brook Sample Archive building after Cindy Veen. Cindy was the Information Manager at Hubbard Brook from 1988 until her untimely passing 1996. She died at the age of 41 after a long battle with breast cancer. Raised in Michigan, she did undergraduate work in geology at Central Michigan University. She earned her master's degree in geophysics from Portland State University and Oregon State University. Before working for the Forest Service in Durham, NH, she was a geophysicist for Mobil in Dallas, Texas. She loved the outdoors, taking every opportunity to hike, bike, kayak, ski and garden.

Cindy was well-respected at the USDA Forest Service and among her peers in the LTER network. Among her many accomplishments, she was instrumental in the establishment of the Hubbard Brook sample archive and worked diligently to make it operational. The archive building now houses approximately 40,000 samples of soil, water, plant tissue, and other materials. Samples are preserved, barcoded, and cataloged with accompanying metadata in a database. Requests for reanalysis of these samples are received periodically, and have resulted in a number of publications.

On the tenth anniversary of Cindy's death, the Hubbard Brook sample archive building was dedicated to her at the Annual Hubbard Brook Cooperatorsâ€™ meeting. Cindy's former supervisor, Tony Federer (US Forest Service, retired), shared some some stories and presented a plaque. The plaque is now on the inside of the archive building and a sign bearing her name is on the exterior, ensuring that her contribution to the Hubbard Brook Ecosystem Study will be remembered.

Controlled-Vocabulary Working Group Report

- John H. Porter (VRC)

The controlled-vocabulary working group conducted a session including all LTER IM's and representatives from NCEAS and ILTER at the 2006 Colorado Information Manager's Meeting. The working session included a brief description of the activities of the working group since the 2005 meeting. These are summarized in a PowerPoint presentation at: http://gce-lter.marsci.uga.edu/lter_im/2006/app/uploads/workgroup1/Controlled_VocabularyIM06.doc and in the Spring 2006 Databits (http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/06spring/#4fa). These efforts had been aimed not at developing a new controlled vocabulary, thesaurus or ontology, but on providing resources to help evaluate the utility in the LTER context of existing controlled vocabularies, thesauri and ontologies such as GCMD, NBII, GEMET and WORDNET resources. The planned aim of the working group was to discuss next steps in terms of the evaluation of

these resources and discussion of tools for exploiting them to aid in LTER data discovery, and two separate groups focusing on these issues were to be established. However, following the presentations, the IM group decided that a fuller reconsideration was needed, so four concurrent working groups simultaneously addressed the same three questions:

1. What do we want to do? Or more specifically: What capabilities do we want to provide for data discovery and data integration?
2. What existing resources could we use and what can they provide?
3. What do we need to do to develop our own resources?

These working groups then reported back to the entire group. Overall conclusions drawn from those presentations were that:

1. Groups were complementary - each group focused on different aspects of the problem
2. The effort was worthwhile. A high priority was placed on improving the ability of LTER data to be searched and browsed.
3. The challenge is complex. It was suggested that the best route was to start simple, and look at other efforts and how they are addressing the challenge.
4. There was a discussion of how many words are useful. A vocabulary with too few words leads to lack of precision while too many lead to a lack of reliability.
5. We need to involve working scientists in the process, since they will ultimately be the "consumers" of the system.
6. We need to know which search terms researchers are actually using to locate data.
7. Information managers need to be involved in the evaluation of alternatives.
8. The products and tools will evolve - this is not a one shot thing.
9. We need to be aware of the substantial work meeting this challenge will require.
10. We need to have an appropriate balance between top down vs. bottom up.
11. We need to decide how broad (in a disciplinary sense) we need to make our vocabulary.

Based on that discussion, we reached 4 major conclusions/action items:

1. We should enable auditing on Metacat to track requests so that we can see how researchers are attempting to use the existing system.
2. We need additional educational activities to help prepare the LTER IM group to deal with the complexities of thesauri and ontologies.
3. We should continue efforts to examine what others have done and to relate that work to our own efforts, such as taking a closer look at NBII, GCMD etc. are doing/have done
4. Continue existing work on compiling lists of attributes and work with the work with SEEK KR group to represent attributes in ontology template.

NSF Workshop: History and Theory of Infrastructure. Lessons for New

Scientific Cyberinfrastructures

- Florence Millerand (PAL/CCE)

An NSF-sponsored workshop on the history and theory of infrastructure was held last September at University of Michigan, School of Information, to address an important question: What practical lessons can the history, sociology, and experience of existing infrastructures offer to the imagination, implementation, and governance of cyberinfrastructure?

The workshop brought together historians, social scientists, and infrastructural practitioners, all engaged on efforts to imagine, develop, and build new cyberinfrastructures. Members of the Comparative Interoperability Study Team (Baker et al, Network News Spring 2005, p17, http://www.lternet.edu/news/images/spring05/NetworkNewsSpring05.pdf) were associated with this workshop as co-organizers, presenters and/or participants. By generalizing the lessons of social and historical analysis, this workshop intended to contribute to the development of infrastructure studies as a distinctive and practically engaged field of study, while also reporting back to the NSF with useful input for the cyberinfrastructure design process.

The topics addressed were organized around four main clusters: Design, dynamics, tensions and strategies. At the design level, optimizing collaboration between social research and the design community was discussed though alternative modes of social scientists' engagements in infrastructure projects - including reading groups and co-design activities with the information management community. Also, participants discussed issues of strategic standardization, flexibility, and sustainability as well as patterns and principles in infrastructure development, tensions of scale in data infrastructure, and organizational capacities.

More information about the workshop: http://www.si.umich.edu/InfrastructureWorkshop/
Texts ("thought pieces") are available at the workshop's blog: http://icd.si.umich.edu/~cknobel/

Good Reads

Ajax And PHP: Building Responsive Web Applications

- Sabine Grabner (MCR)

C. Darie, B. Brinzarea, F. Chereches-Tosa, M. Bucica. Ajax And Php: Building Responsive Web Applications. Packt Publishing, 2006

If you thought lately that your web applications need more life, you probably want to know more about AJAX (Asynchronous JavaScript and XML). In a nutshell, it is the use of the nonstandard XMLHTTPRequest() object to communicate with server-side scripts. It can send, as well as receive, information in a variety of formats including XML, HTTP, and even text files. Ajax's most appealing characteristic, however, is its "asynchronous" nature, which means it can do all of this without having to refresh the page. This allows you to update portions of a page based upon user events [Web Design In A Nutshell, 3rd Edition, Jennifer Niederst Robbins, 2006].

AJAX and PHP is a great resource for getting started with AJAX, as well as web development. By introducing the technique of AJAX in the first chapter, followed by easy reading theory and several chapters of different applications including very useful code, the book lays out the basic architecture of server and client side web programming and gives enough information to allow a newbie with background in object-oriented programming (OOP) to get started or for the advanced web developer to jazz up existing applications.

At this point I spare to list the content of the book, as it is very well documented at the books webpage, where you can find two chapters in pdf and all the source code. I also want to refer to author Cristian Darie's comment on the book at amazon.com.

By the way, since the backbone of AJAX is the way you write JavaScript, don't worry too much about the PHP part, it is a nice way to learn it by example.

I did not publish any link to my AJAX sites here because they are still subject to change, but feel free to contact me to look at the pages.

Metadata: Implementation of an International Framework

- Karen Baker (PAL)

C-C Lin, J.H.Porter and S-S Lu, 2006. A Metadata-based Framework for Multilingual Ecological Information Management. Taiwan Journal of Forest Science 21(3): 377-382.

In this paper, an ILTER team has summarized a collaborative effort. It provides an overview of metadata in general and of an example of deployment of the Ecological Metadata Language in particular. The authors identify metadata as a "critical tool" for data collection when there is an intention to make data available over the long-term. In describing a metadata-based framework, a brief history of metadata and of LTER publications on this topic is provided. The framework used is organized into a three tier system: data storage and management, data editing and processing, and, finally, user interface. Some project statistics of implementation are reported. Such a summative report contributes to a contemporary need for effective assessment mechanisms in the field of information systems. In considering the "problem of integrating ecological data", the authors provide an important, thought-provoking prompt for data providers as well as for information managers. There is a lot of information pertinent to most any data preservation effort packed into this metadata paper.

Data Curation in E-Science

- Ted Gragson (CWT)

KARASTI, Helena, Karen S. BAKER, & Eija HALKOLA. 2006. Enriching the Notion of Data Curation in E-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network. Computer Supported Cooperative Work 15(4): 321-358.

This article provides an ethnographic analysis of ecological data stewardship in the LTER - arguably one of the most informative long-run experiment in consistently collecting, managing and providing data to an ever-evolving scientific community of practice. While the origin of the LTER over 25 years ago was visionary, trial-and-error has helped ensure the vision did not devolve into ideology. A contemporary vision, e-Science or Cyberinfrastructure, is touted as the platform on which scientific research and education in our emerging knowledge economy is to be built. The vision is one of data

generation and sharing by way of technologies neither previously available nor imagined by scientists and their students. This article focuses on the practices of long-term ecological data stewardship to draw attention to the connection between means and ends; this informs the more technological, sometimes ideological, discussion about data grids, computing power, middleware, and automation characteristic of e-Science. As such, this article offers practical information that will help bring closer to reality the vision of e-Science and the growing necessity of collaborative partnerships among diverse sciences to both understand environmental change and inform future decision-making.