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

01001100 01010100 01000101 01010010

01001100 01010100 01000101 01010010

◆ **Feature Articles**

About this Issue

New Information Management System tools developed by the Florida Coastal Everglades (FCE) LTER Program

Kepler: A System for Scientific Workflows

Introducing WatershedDB

EML Harvesting I: Metacat Harvester Overview and Management

◆ **News Bits**

Forest Service Research Station Launches New Database

Ecological Metadata Language in Brazil and Costa Rica

◆ **Good Reads**

Good Read: Infrastructuring for the Long-Term: Ecological Information Management - John Campbell (HBR)

The Cognitive Style of PowerPoint - Karen Baker (PAL/CCE), Jerry Wanetick (CCE/PAL), Shaun Haber (PAL/CCE)

Data at Work: Supporting Sharing in Science and Engineering - Karen Baker (PAL)

DataBits: An electronic newsletter for Information Managers ----- Fall 2004 Issue

Featured in this issue:

- New information management tools for
    - Interactive mapping,
    - Data management,
    - XML conversion,
    - Scientific workflows
    - Data Harvesting
- ClimDB and HydroDB go spatial with the formation of WatershedDB
- News Bits
- Good Reads

DataBits is 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.

◆ Feature Articles

New Information Management System tools developed by the Florida Coastal Everglades (FCE) LTER Program

- Linda Powell, Florida Coastal Everglades LTER

Information and Project Management Tool

The FCE LTER has added a new web-based interactive mapping application called the 'FCE LTER Interactive Everglades Map' to their website at http://fcelter.fiu.edu/gis/everglades-map/ to facilitate information and project management. Project and data management at the FCE LTER are very complex because of the large number of research personnel (114) and immense study area. The majority of Florida Coastal Everglades LTER sites are located in freshwater marsh, estuarine mangroves, and seagrass estuary ecosystems in Everglades National Park, an area that covers approximately of 4300 km2 or 1,062,553 acres. With 21 'official' FCE LTER and 260 'related' sampling sites, the task of keeping track of the different types of research, sampling parameters, sampling frequencies, and intra-site sampling point locations posed to be very difficult. Additionally, the scientists studying the Florida Everglades have generated numerous amounts of historical data over the past 6 decades, some of which the FCE LTER is looking to archive in their information management system.

The FCE information management team of Linda Powell (FCE Information Manager) and most notably Mike Rugge (FCE Program Manager) have collaborated to design the 'FCE LTER Interactive Everglades Map' tool that can deliver FCE program information via the web. There are two main components of the application: 1) Interactive Mapping and 2) Project and Data Management.

Interactive Mapping

Figure 1 - A screenshot of the Florida Coastal Everglades interactive Everglades Map

The interactive mapping feature allows the web user to build maps to their specifications by toggling GIS layers such as FCE sampling sites, major roads, major canals, and boundary outlines. These layers of information can then be overlaid onto Landsat7 ETM imagery, land use, or digital orthoquad (DOQ) backgrounds. Users have the option to either save or print their map creation. All the GIS layers displayed in the interactive map are available for immediate download to the user.

Data and Project Management

A web user also has the ability to display details about a particular project or site by graphically selecting an area of interest from the map or directly choosing a site from a drop-down menu. When the user selects a 'Project' from the menu, a project description page offers the project's name, contact information, start and end dates, related abstract. From this page, the user can choose to display a list of all sites, sampling attributes, publications, and personnel for the project of interest. Filters allow the user to narrow the list of sampling attributes by workgroup and keywords or the publications by author's name and keywords. The 'Site' selection displays a page with the site's name, affiliated projects, alias, latitude/longitude, size, and descriptions of its watershed, hydrography, topology, geology, soil,

vegetation, habitat, and climatology. Here a user can choose to display a list of sampling attributes, sampling attributes by dataset, datasets, and publications related to this particular site. If the user chooses to view all datasets affiliated with the site, the list contains active links to dataset metadata and downloads.

Our researchers can use this application to help with future experimental designs, publication discovery, and intra-site syntheses. The information management team uses the tool to facilitate data management as they can easily see a list of sampling attributes for each site and a list of sampling attributes by datasets for comparison. If there is not a cross-reference between the lists, the information manager knows which data have not been submitted, which project is responsible for generating those data, and whom to contact regarding data submission.

We used the following interactive mapping software and tools:

1) Mapserver - Developed by the University of Minnesota ForNet project and enhanced by the Minnesota Department of Natural Resources (MNDNR), open source
2) PHP MapScript - Developed by DM Solutions Group, open source
3) ROSA Java Applet - Developed by DM Solutions Group, open source
4) GMap mapping engine (PHP/MapScript version) - Developed by DM Solutions Group, open source.

EML Excel Metadata Template and XML Converter Tool

Over the past few years, many institutions within the ecological community have been working together in order to create a common syntax for sharing ecological data and metadata. The effort has produced the Ecological Metadata Language (EML) (http://knb.ecoinformatics.org/software/eml/), which is a formalized specification for expressing information about ecological resources. The language that EML itself is expressed in is the Extensible Markup Language (XML), which is an open, Internet-based standard that was designed to enable the creation of discipline-specific languages such as EML. The Long-Term Ecological Research (LTER) community has adopted EML as its metadata content and format standard and sites within the LTER network are now working to retrofit existing metadata into EML, as well as creating EML for the newly submitted data.

Since the inception of the Florida Coastal Everglades (FCE) LTER Program in May of 2000, FCE researchers have been required to submit an Excel metadata file with their data; otherwise, the data would not be accepted by the FCE Information Manager. One of our biggest challenges in the EML implementation process was finding a way that FCE researchers could continue to collect metadata using an Excel metadata template AND easily produce EML (XML) documents. The Excel format facilitates metadata collection as all of the researchers are very familiar with the spreadsheet environment. Recently, Mike Rugge, the FCE program manager, finished developing a Perl application that will convert the FCE Excel metadata into a valid EML document.
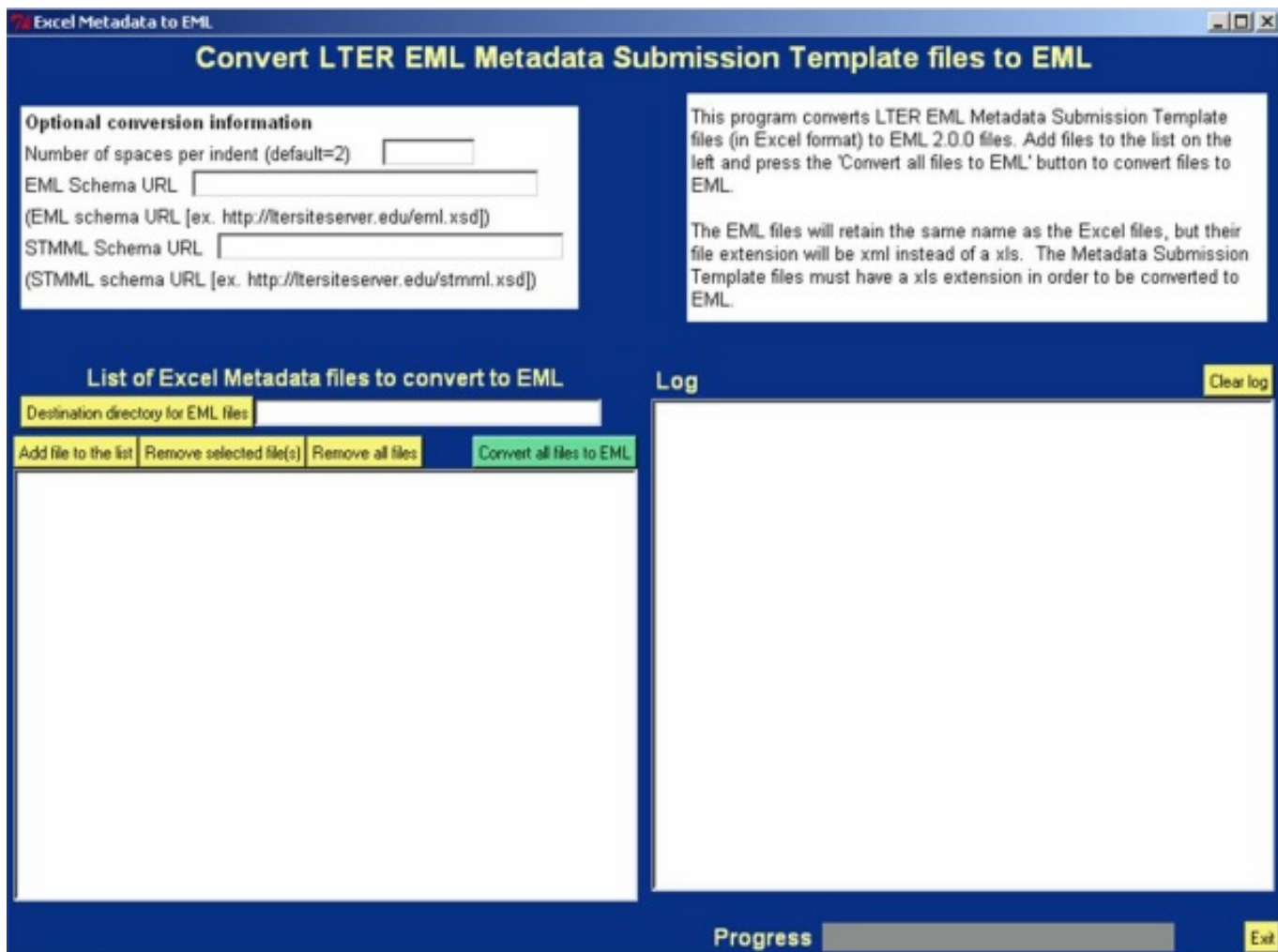
Figure 2 - A screenshot of the Florida Coastal Everglades PERL application

The application allows the user to process one or many Excel (.xls) files in one run and to choose the destination directory for the newly created EML documents. Minor adjustments to the structure and content of the original FCE Excel metadata template were made to explicitly follow EML and to better address the issue of possible multiple entries, such as multiple dataset creators or keywords. Linda Powell, FCE Information Manager, demonstrated this tool at the annual LTER Information Managers' Meeting in Portland, Oregon this past July and has made the Excel metadata template and converter tool available to the LTER network, with the hope that other sites may use them as an aid to their EML implementation work.

Figure 3 - A screenshot of the Florida Coastal Everglades metadata template

Those interested in a copy of the EML Excel Metadata Template and XML Converter Tool may contact Linda Powell at powell@fiu.edu. In the near future, the template and converter tool will be available via download from the Information Managers' Mentor web page hosted by the LTER Network Office (LNO).

Kepler: A System for Scientific Workflows

- Dan Higgins (higgins@nceas.ucsb.edu) and Matthew B. Jones (jones@nceas.ucsb.edu) National Center for Ecological Analysis and Synthesis, University of California Santa Barbara

Scientists in a variety of disciplines (e.g., biology, ecology, astronomy) need access to scientific data and flexible means for executing complex analyses on those data. Such analyses can often be described in terms of a number of distinct operations with the results of one operation being passed to the next. This overall process describing data flow from one operation to the next can be called a 'scientific workflow'. A formal description of this workflow allows for efficient execution and repetition of such analyses, as well as providing documentation of exactly how data were analyzed. Kepler is designed to aid scientists in the design, construction, execution, and communication of such scientific workflows. For example, Kepler uses structured metadata such as Ecological Matadata Language (EML) metadata to make it easy for scientists to locate, analyze, and visualize unfamiliar data from data repositories around the world. Kepler includes a tool for creating graphical displays of workflows in the form of 'boxes', which represent operations or analytic steps, connected by 'arrows', which indicate the flow of information between the workflow steps. Kepler includes flexible mechanisms for controlling the data flow or sequencing the operations in these workflows, and for executing, saving, and re-creating such workflow descriptions.

Kepler is currently under development by the Kepler Project, a collaboration of various projects to develop open source tools for scientific workflows. These contributing projects include SEEK (see "Building SEEK: the Science Environment for Ecological Knowledge" by William Michener in an earlier DataBits issue), SDM/SPA, PtolemyII, GEON, and

ROADNet (see the [Kepler web](#) site for further information). It is important to note that Kepler is based on the [Ptolemy II](#) project. Ptolemy is a modeling, simulation, and design effort that has been going on for more than ten years at the Department of Electrical Engineering and Computer Science (EECS) at UC Berkeley. The baseline Ptolemy II software was first created over 5 years ago, and is thus tested, stable and well documented. Building Kepler on this existing base avoids having to reimplement a large amount of software and allows the effort to concentrate on new features.

In Ptolemy II nomenclature, the 'boxes' in a graphical workflow are called "actors". Each actor may have "input ports" and "output ports" which are connected by 'arrows' to other ports. The ports and their connections represent paths through which data move in the workflow. A Ptolemy II model also has a "director" which coordinates the actions of the "actors". Perhaps the simplest type of director just tells actors to wait until data appears at an input port to start processing the data and then transfer it to an output port. When that data appears at the output port, it flows to whatever inputs it is connected to. Although this simple dataflow based on availability is often appropriate, one of the great features of Ptolemy II is that the director can be changed to allow for other types of models. For example, actors might be directed to 'fire' at prescribed, simulated times, whether or not input data is available.

The "actor" in Kepler can be thought of as a software component that processes the data that appears at its input ports. In addition to the ports, actors can also have "parameters" which the workflow designer sets to control just what the actor does. Actors can be very simple; for example, one may take an array of numbers as an input and simply count the number of items in the array. Or actors can carry out very complex operations like a genetic algorithm predicting species abundances based on a variety of environmental factors. Composite actors are also allowed, where a complex workflow is 'hidden' within a single actor 'box'. This allows for hierarchial workflows where certain complexities are 'hidden' to help understanding. There are also 'source' actors, which provide data and do not require inputs, and 'sink' actors which usually are just displays of data and have no outputs.

Actors which are available for building a Kepler workflow appear on the left of the graphical tool display, as indicated in Figure 1. Ptolemy provides about 100 actors and, so far, the Kepler program has added roughly 100 more. Workflows are built by dragging actors from the left onto the panel on the right and then connecting them to represent the workflow being created. Users can also build their own actors, either by programming in Java, configuring scriptable actors (curerntly by using Matlab, R, or Python scripts), or by building composite actors from low level actors.
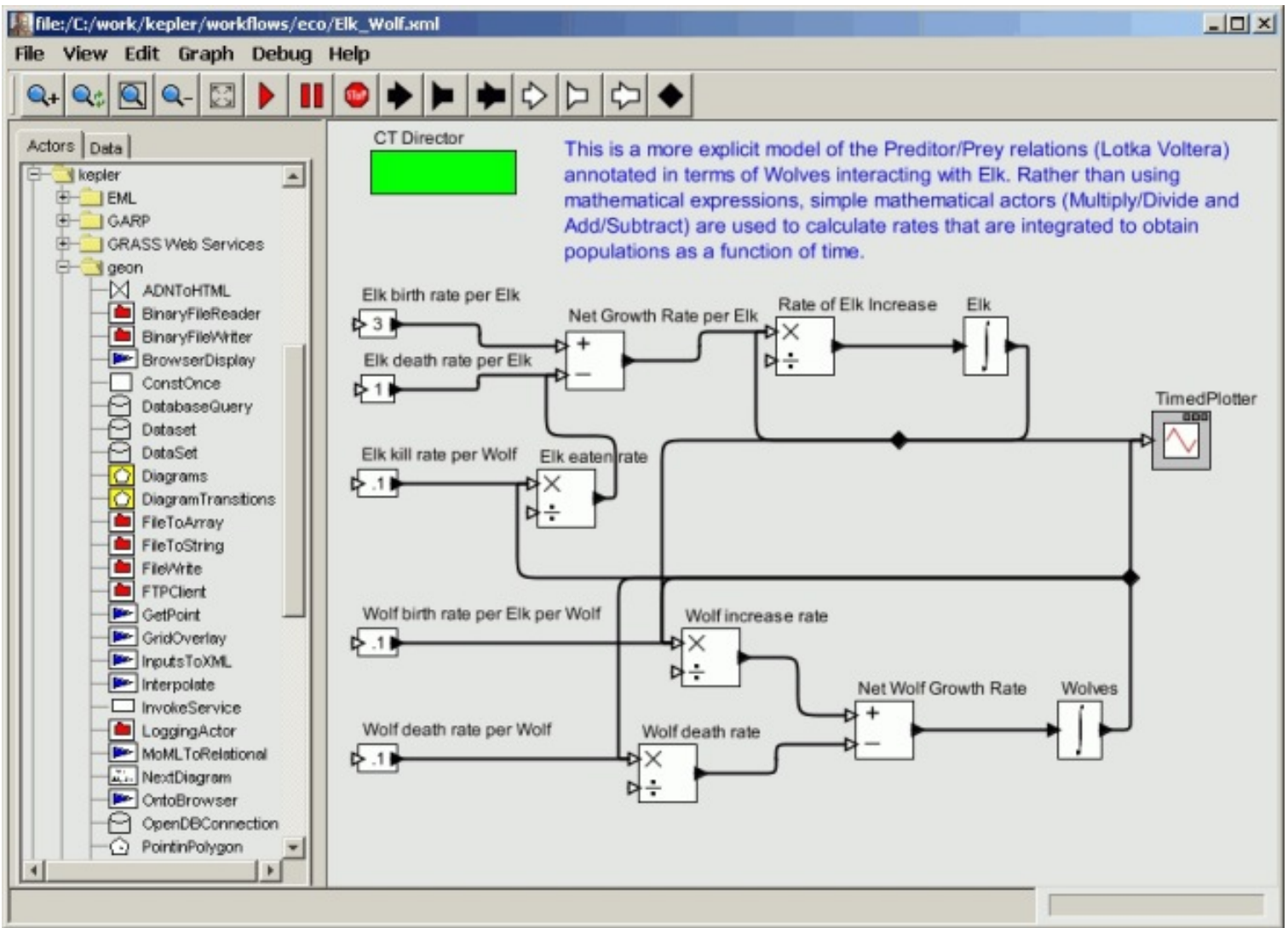
Figure 1 - A screenshot of Kepler showing the "Actor" tab and a Predator/Prey Workflow

One of the efforts in Kepler is to provide a variety of specialized actors and help scientists locate and use computing services that have been created elsewhere. As an example, actors for accessing web services have been created, as have actors for accessing Ecogrid data (see "SEEK Ecogrid" article in the Spring 2003 issue of "SEEK Ecogrid" article in the Spring 2003 issue of Databits). The web services actors can be used, for example, to carry out bioinformatics analyses on a web server in Japan or a geospatial image processing service in California and then automatically pass the results to a local actor for further processing. A data streaming example has also been created which displays almost real-time images from a remote location (Figure 2).
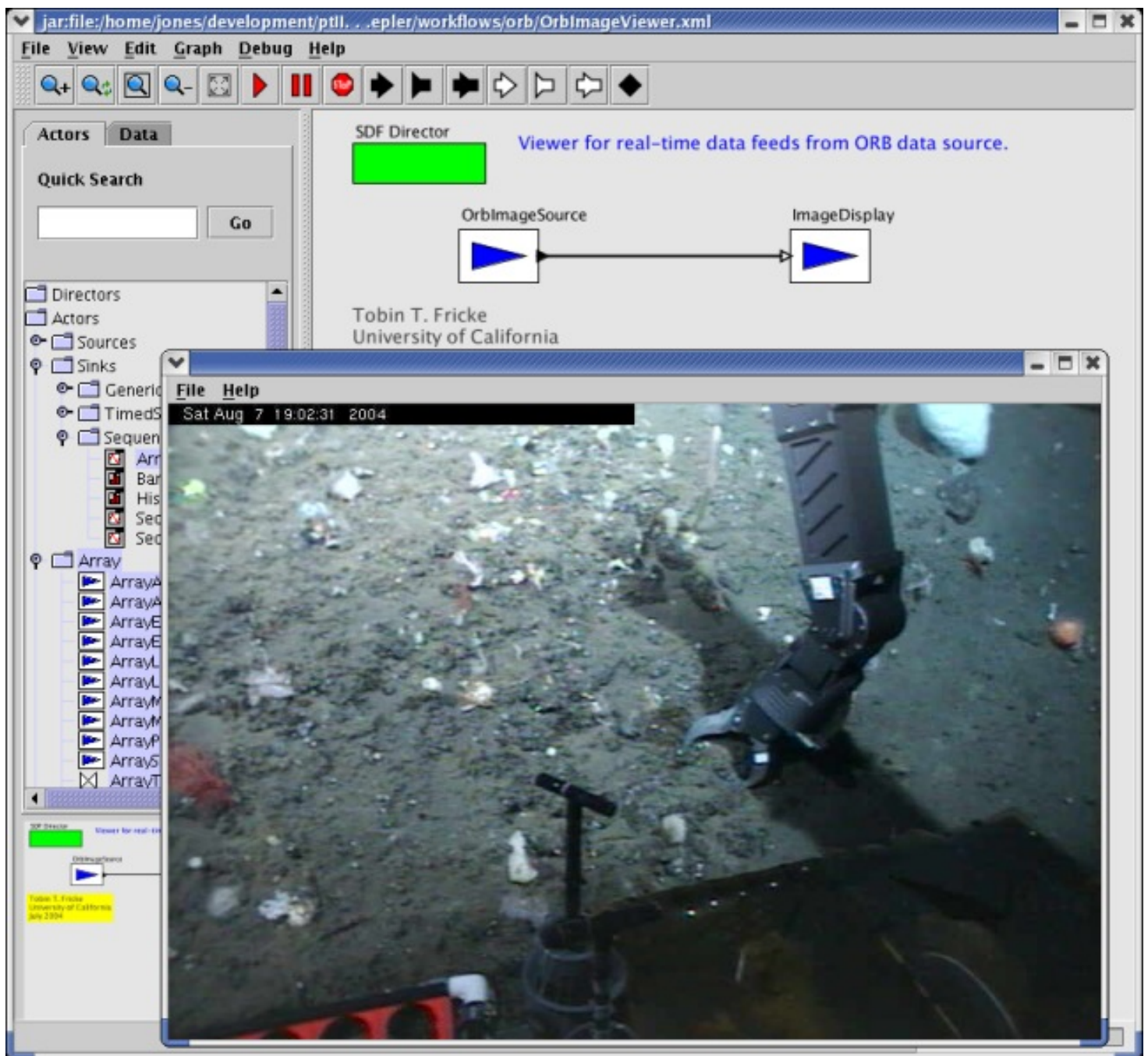
Figure 2 - A Kepler screenshot showing the deep-sea floor from a real-time data source from a submersible. Various signal and image processing utilities are available to analyze as well as display temporal data streams such as this video stream.

A final relevant example is illustrated in Figure 3. On the left of this screen, the "Data" tab has been selected and a search has been carried out for data sets on the Ecogrid. One of the resulting datapackages has been 'dragged' onto the graphics display on the right. This action results in the creation of the EMLDataSource actor shown in the figure. The EML metadata is automatically used to determine the number of columns and datatypes of the data described by the EML document. In this case, there are 11 columns in the data table and an output port (one of the black, right pointing triangles) is created for each column. Two of these ports are connected to a plotting actor (labeled "HumidityPressurePlot"), and when the workflow is executed, the actual data is retrieved from the Ecogrid and plotted in the window shown in Figure 3. This example shows how Kepler can help run analyses on remote data by leveraging EML and the Ecogrid, making it an excellent platform for exploratory analysis and visualization of unfamiliar data.
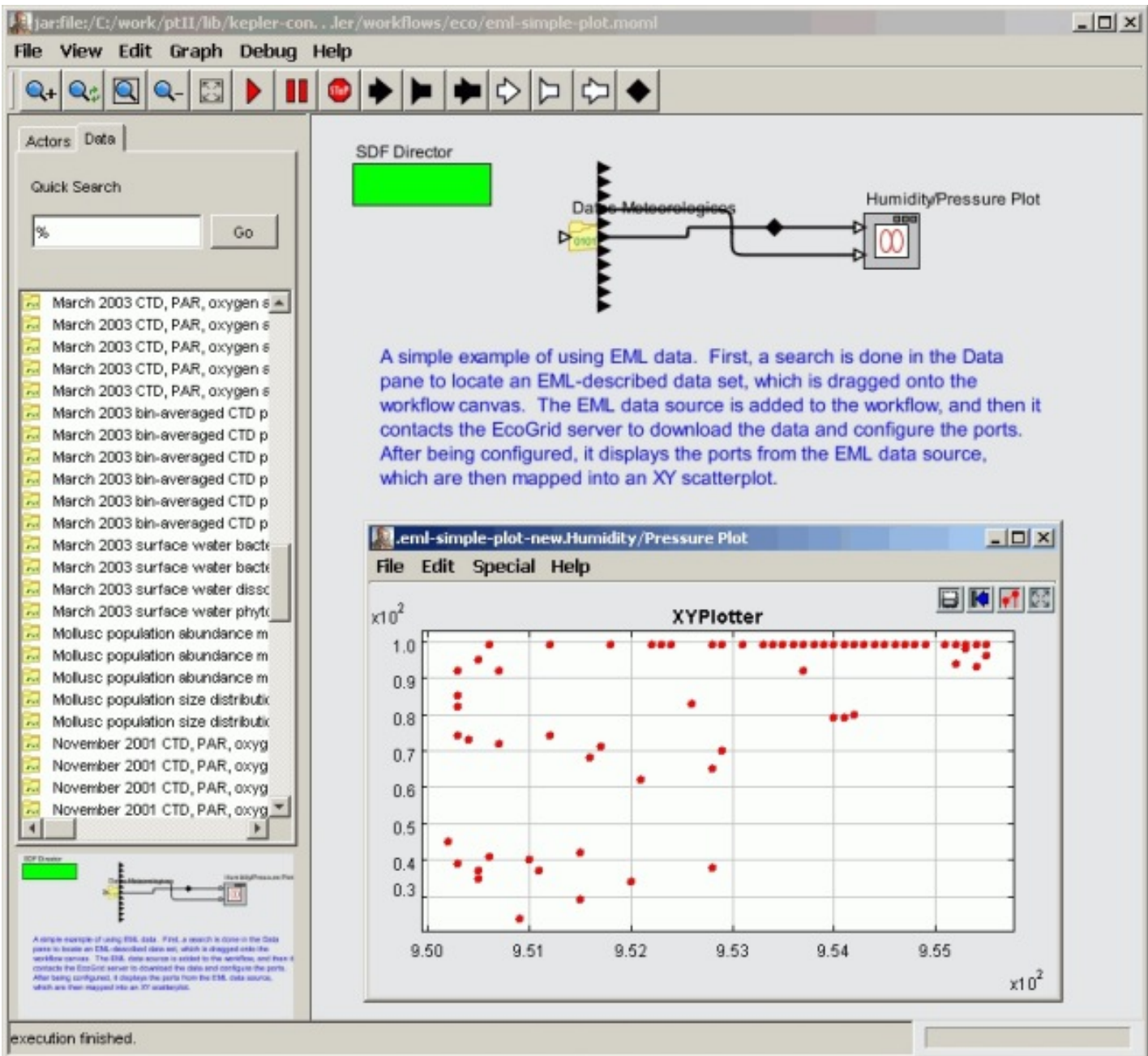
Figure 3 - A Kepler screenshot showing Ecogrid data search and EMLDataSource results. Users can search for data that is available on the EcoGrid and use it directly in workflows as if it resided locally on their computer.

Due to its brevity, this article can only provide a brief glimpse of Kepler and its possibilities. For example, there is an effort to add semantic information in the form of ontologies to data searches and workflows in order to facilitate more automatic processing, error checking, and data discovery. Also, the Kepler project is on-going and thus continually changing. Additional capabilities are being added all the time. The interested reader should visit the Kepler website for more information and a view of the current state. Comments and suggestions are always appreciated, especially from potential users, and we welcome new contributors to participate in the design and development of the Kepler system.

Acknowledgements

Introducing WatershedDB

- Theresa Valentine, Andrews LTER

WatershedDB is a project to assemble spatial data for participating experimental watersheds in the USFS/LTER-sponsored ClimDB/HydroDB project. The goal is to build an Internet accessible application to allow researchers to view and interact with site spatial data layers, and also to link to the data collected and stored in HydroDB. We will also investigate working with ClimDB sites. Funding has been provided by the USFS WO-R&D to begin work on this project during the summer of 2004.

The purpose of this article is to introduce WatershedDB to the LTER community, and get feedback on our plans. At this time, we have asked Forest Service sites to provide us with data. We will attempt to gather information from LTER sites through their data access web pages. We will then be contacting individual LTER sites if we need additional data. Our first priority is to obtain data from Forest Service sites, as they are providing the funding for the project. Please submit any comments/concerns to theresa.valentine@orst.edu

Priority datasets for the project include:

a. The Experimental Forest Boundary (if applicable)
b. The context watershed boundary (this might be the Experimental Forest boundary or the larger watershed boundary that includes all of the gauged watersheds)
c. Individual gauged watershed boundaries (actual shape, not bounding coordinates)
d. Gauging Station locations (lat/lon in decimal degrees)
e. Meteorological Station locations (lat/lon in decimal degrees)

   Other data of interest include:

f. Stream network
g. Digital Elevation Model for the context area
h. Roads
i. Other GIS data available for the experimental watersheds

Additional data for the entire area will be collected and used as backdrops for the Internet mapping application (temperature, precipitation, ecoregions, political boundaries, etc).

The first product will be a pilot study that will look at two options for collecting, storing, and serving data over the Internet. The first option is to collect the data from all the sites, combine it into an SDE Geodatabase and serve the data from one site (consolidated approach). The second option is to gather data from sites that don't have Internet mapping capability, combine it into one SDE Geodatabase, and connect these data with SDE Geodatabases located at individual sites (distributed approach).

The initial plans are to develop feature classes for priority datasets (a-e). We will be creating the feature classes, standardizing the field names and definitions (checking with HydroDB and ClimDB standards), and loading the data into the Geodatabase. This project will be using data from sites that have agreed to participate. The data for the HJ Andrews Experimental Forest already resides in an SDE Geodatabase. These data will be used to test the distributed approach.

After data has been converted into SDE, an ArcIMS site will be built to allow users to interact with the data. Priority functions will include:

- Turning layers on and off

- Zooming in for more detail (scale dependent)
- Identifying features
- Querying features
- Querying databases

HTML and Java Internet Mapping applications will be built and compared.

EML Harvesting I: Metacat Harvester Overview and Management

- - Duane Costa (LNO)

Introduction

With the number of EML datasets produced at LTER sites growing steadily, a need arose for an automated mechanism to upload large numbers of EML documents from various sites to a central Metacat repository. The Metacat Harvester software, recently developed at the LNO as part of the Knowledge Network for Biocomplexity (KNB) effort, fills such a purpose. It provides batch upload of EML documents from one or more sites to a Metacat server on a regularly scheduled basis.

Prior to the development of the Metacat Harvester, there were two ways that EML documents could be uploaded to Metacat. A document could be submitted to Metacat via the Morpho tool, or it could be submitted through a Metacat web interface. In both cases, the document was pushed from the client side to the Metacat server, the upload was manually initiated by the client, and documents could only be uploaded one at a time.

The Metacat Harvester adds a third important way to upload documents to Metacat. In contrast to the client-initiated push of Morpho or the Metacat web interface, the Metacat Harvester automates the process by using a regularly scheduled server-initiated pull. In addition, the Metacat Harvester can retrieve large numbers of documents from multiple sites in a single harvest, dramatically increasing the rate at which new EML documents can be added to the LNO Metacat metadata catalog. EML documents uploaded to the LNO Metacat are also automatically replicated to the KNB Metacat (knb.ecoinformatics.org) metadata catalog, where they may be accessible to an even wider audience.

Harvester Features

The Metacat Harvester software, a Java application, is a component of the upcoming Metacat 1.4.0 release. Although Harvester is bundled with the Metacat distribution, it runs as an independent application. Harvester is supported on the Windows, Unix, and Linux platforms. The Harvester software implements the following features:

- Each site that registers with Harvester controls its own harvest schedule. A site can schedule harvests as frequently as once per day, or as infrequently as once every 6 months or longer.
- Harvester will not try to reharvest EML documents that are already present in the Metacat repository. Harvester will only reharvest a document if the revision number of the site's document is higher (more current) than the revision number of the document in the Metacat repository.
- Harvester generates and sends an email report to the site after every harvest. The email address that the report is sent to is determined during the Harvester registration process.
- Harvester logs all of its operations in the same database that Metacat uses. This means that the individual who administers Metacat can inspect the database to view records of all the operations that Harvester has performed.
- Harvester works with dynamically generated EML. Harvester can retrieve an EML document from any valid URL regardless of whether the URL returns a static XML file or generates its XML dynamically.

Using the Metacat Harvester

There are two roles to describe with regard to using Harvester. The first is the role of the *Harvester Administrator*, the

individual who installs and manages the Harvester software. Typically, this is the same person who installs and manages a Metacat server. The Harvester Administrator performs the following actions on the server side:

- configures the Harvester software;
- runs the Harvester application;
- reviews email reports sent to the Harvester Administrator after a harvest.

A second role is that of the *Site Contact*. This may be a LTER Information Manager who is interested in registering his or her site to schedule regular harvests. The Site Contact performs the following actions on the client side:

- composes a *Harvest List*, an XML document that describes the names of the documents to be harvested, the version of EML the documents conform to, and the locations (URLs) from which the documents can be retrieved;
- registers with Harvester by logging in to a web form and entering a few items of information, such as the URL of the Harvest List;
- reviews email reports sent to the site after a harvest.

The role of the Site Contact is more fully described in part two of this article, EML Harvesting II -- Preparing Site Metadata and Harvest Lists. Additional information is available [online](#) on the IM Mentoring page at the LTER web site.

Who's Using Harvester?

To date, only the LNO has deployed Metacat Harvester, though it will soon be available for more general use in conjunction with the pending release of Metacat 1.4.0. Four LTER sites have registered to schedule regular harvests with the LNO Harvester, ranging in frequency from one harvest per week to one harvest per month, and several more sites are expected to register soon. Approximately 300 new EML documents have been harvested to the LNO Metacat thus far, *and the number continues to grow!*

◆ News Bits

Forest Service Research Station Launches New Database

- Jonathan Walsh - BES

The Forest Service has announced Treesearch, a new database that allows access to Forest Service research publications online. With Treesearch, customers can locate and download agency-authored or sponsored publications, including those in journals, books, and conference proceedings. The research results have been peer-reviewed to ensure the highest quality science. The publications in Treesearch can be accessed from www.treesearch.fs.fed.us. The database supports searches by author, keyword, originating organization, or date; and returns the abstract and, if selected, the full text of the publication, including tables, figures, and citations. *This announcement was originally posted in the USDA Forest Service July Urban Projects newsletter. (http://www.fs.fed.us/na/morgantown/uf/news/July2004news.pdf*

Ecological Metadata Language in Brazil and Costa Rica

- From 2004 ILTER Coordinating Committee Meeting Proceedings (http://www.ilternet.edu/)

Two presentations of special interest to Information Managers were given at this year's ILTER CC Meeting - one for the Brazilian "PELD" LTER network and their "digital library" using open source tools and Ecological Metadata Language (EML). Also a presentation on the "TEAM" project which also uses EML.

◆ Good Reads

Infrastructuring for the Long-Term: Ecological Information Management

- John Campbell (HBR)

Helena Karasti and Karen S. Baker, "Infrastructuring for the Long-Term: Ecological Information Management", in Proceedings of the 37th Hawaii International Conference on System Sciences, Big Island, Hawaii, January 5-8, 2004.

(http://csdl.computer.org/comp/proceedings/hicss/2004/2056/01/205610020cabs)

This paper emphasizes the critical role of information management in designing collaborative long-term ecological research. The LTER Network is used as a case study to examine information management within the context of large-scale scientific research. The authors' analyses draw largely from interviews with LTER scientists, information managers and others involved in the LTER program, using quotations and anecdotes to highlight points. A description of how information management provides support to science, data, and technology is provided. While information management is typically carried out behind the scenes, the authors make a convincing argument that it is a fundamentally important part of the infrastructure of collaborative research. This paper offers an inside look into the workings of information management within the LTER program and is a good read for information managers, scientists and others involved in long-term ecological research.

The Cognitive Style of PowerPoint

- Karen Baker (PAL/CCE), Jerry Wanetick (CCE/PAL), Shaun Haber (PAL/CCE)

Tufte, E.R., 2003. The Cognitive Style of PowerPoint. The Graphics Press, Connecticut 28pp. (www.edwardtufte.com)

Tufte's work is concerned with careful awareness in the presentation of quantitative information. His books "Envisioning Information" and "The Visual Display of Quantitative Information" are recognized as works of art and of insight. The "Cognitive Style of PowerPoint" represents a full-blown PowerPoint rant making clear some of the dangers inherent in any well-packed slide presentation. Tufte discusses the over-simplification of data, the burying of information in deeply nested bullet point levels, and the consequences of omitting context which allows an audience to connect the flow of information between slides. Perhaps this slim phamphlet is a good reminder for LTER Information Managers of just why their recent meeting in Portland was designed to avoid the sit-and-receive style of PowerPoint communication,

providing instead a multi-method, multi-voice participatory approach more apt to result in dialogue representing a diversity of views.

Online 'Wired' magazine creates a powerful message about technology in general and PowerPoint in particular through the juxtaposition of two perspectives: the danger (Edward Tufte; PowerPoint Is Evil; http://www.wired.com/wired/archive/11.09/ppt2.html) and the potential (David Byrne; Learning to Love PowerPoint; http://www.wired.com/wired/archive/11.09/ppt1.html). Musician and artist David Byrne sees PowerPoint from an artistic perspective. While he agrees that PowerPoint often produces cheap-looking slides that poorly display content, he argues that it also holds the potential for creating artistic content-free slides, which may complement a presentation. Employing various types of media (photos, graphics, movies, music, etc.), a user may create visually-stunning yet cohorent slides, even without the need for words. Byrne considers the medium itself to be the content. He is aware that this software is "limiting, inflexible, and biased..." but finds this "a small price to pay for ease and utility" of outlining a presentation, especially when it also represents an artistic outlet, useful for creating visual aesthetics that convey thought and emotion.

These dual perspectives are reflected in the name itself, 'PowerPoint': on the one hand, 'power' as in dominate the audience and 'point' as in heirarchical bullets driving home a pitch; on the other hand, 'power' as in organized, synthesized information and 'point' as in another point of view. Tufte reminds us that the amount of packed information is not automatically related to what the speaker understands and is rarely correlated to what the audience will comprehend. Yet it's not the technology but rather the time-short, product-packed speaker that decides whether PowerPoint will be used as a "talk substitute" or a "talk supplement". Tufte's phamphlet is a timely reminder that PowerPoint slide presentation software provides us a tool to be used or misused as one of the steps in organizing, synthesizing, articulating and communicating information.

Data at Work: Supporting Sharing in Science and Engineering

- Karen Baker (PAL)

Birnholtz, J. And M. Biertz, 2003. Data at Work: Supporting Sharing in Science and Engineering. Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work (GROUP'03; 2003 November 9-12). E.M.Tremaine (ed). ACM Press 34: 339-348.

Collaboratories are organizational structures that support distributed work, bringing together scientists with tools and information. Often the collaboratory concept is associated with remote use of telescopes or microscopes. This article focuses not on the collaboratory itself nor its tools but rather highlights data sharing as nontrivial ("data sharing is not easy") and as needed ("Funding agencies appear to be convinced that their underlying 'need' for groundbreaking scientific research will be more effectively satisfied if there is more data sharing among scientists.) Many thanks to James Brunt for sharing this paper over dinner one evening at the recent Information Manager meeting as it presents one aspect of the multifaceted work being done by a team of researchers studying collaboratories at the University of Michigan School of Information.

Categorization of data routinely opens up dialogue within an information management community; this paper adds to such discussions presenting 'data as events' to contrast with 'data as streams' and considering 'data as science enabler' while recognizing research areas that have 'low task uncertainty and high mutual dependence' in contrast with others with 'high task uncertainty'. The notion of needing to understand data practices as one critical element of collaboratories comes as no surprise to an information manager, but the vocabulary and language used to frame the discussion brings valuable definition to some frequently unarticulated thoughts regarding data at work.