

Proceedings of the  
**Environmental  
Information Management  
Conference 2008**  
(EIM 2008)



# Proceedings of the Environmental Information Management Conference 2008 (EIM 2008)

September 10-11, 2008  
Albuquerque, NM

*Editors:*

**Corinna Gries**

*Central Arizona-Phoenix LTER  
Arizona State University*

**Matthew B. Jones**

*National Center for Ecological Analysis and Synthesis (NCEAS)  
UC Santa Barbara*



## Copyright

© 2008

Authors who submitted to this conference agreed to the following terms:

- a) Authors retain copyright over their work
- b) This work is released under a Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0/>), which allows others to freely access, use, and share the work as long as they provide an acknowledgment of the work's authorship and its initial presentation at this conference.
- b) Authors are able to waive the terms of the CC license and enter into separate, additional contractual arrangements for the non-exclusive distribution and subsequent publication of this work (e.g., publish a revised version in a journal, post it to an institutional repository or publish it in a book), with an acknowledgment of its initial presentation at this conference.
- c) In addition, authors are encouraged to post and share their work online (e.g., in institutional repositories or on their website) at any point before and after the conference.

## Table of Contents

<b>Program Committee</b> .....	5
<b>Sponsors</b> .....	5
<b>Program at a glance</b> .....	6
<b>Detailed Program</b> .....	7
<b>Contributed Oral Papers</b> .....	13
Brenda Leroux Babin and Lei Hu .....	14
Scott Bainbridge, Mark Rehbein, Gavin Feather and Damien Eggeling .....	19
Derik Barseghian, Ilkay Altintas, Matthew B. Jones .....	26
B. J. Benson, L. Winslow, P. Arzberger, C. C. Carey, T. Fountain, P. C. Hanson, T. K. Kratz, S. Tilak .....	33
Molly E Brown .....	39
Helen Conover, Kathryn Regner, Manil Maskey, Jessica Lu, Xiang Li, H. Michael Goodman .....	46
Judith B. Cushing, Nicole E. Kaplan, Christine Laney, Juli Mallett, Ken Ramsey, Kristin Vanderbilt, Lee Zeman, Jincheng Gao, Judith Kruger, Carri LeRoy, Daniel Milchunas, Esteban Muldavin .....	52
Donald L. Henshaw, Fred Bierlmaier, Barbara J. Bond, and Kari B. O’Connell .....	62
Lei Hu and Brenda Leroux Babin .....	69
Adam M. Kennedy, Suzanne M. Remillard, Donald L. Henshaw, Lawrence A. Duncan, Barbara J. Bond .....	77
O Langman, PC Hanson, SC Carpenter, K Chiu, YH Hu .....	83
Ben Leinfelder, Jing Tao, Duane Costa, Matthew B. Jones, Mark Servilla, Margaret O'Brien, Chad Burt .....	90
Yong Liu, David J. Hill, Tarek Abdelzaher , Jin Heo, Jaesik Choi, Barbara Minsker, David Fazio .....	96
Shelly L. Moore, Shelly Walther, and Larry D. Cooper .....	105
Margaret O'Brien and Shannon Harrer .....	111
Giriprakash Palanisamy, Bruce Wilson, Ranjeet Devarakonda, Jim Green .....	119
John Porter and David E. Smith .....	125
Ian Sears .....	133
Mark Servilla, Duane Costa, Christine Laney, Inigo San Gil, and James Brunt .....	139
Wade M. Sheldon .....	145
John Vande Castle and Mark Servilla .....	151
Kristin L. Vanderbilt, David Blankman, Xuebing Guo, Honglin He, Jianhui Li, Chau-Chin Lin, Sheng-Shan Lu, Burke Chih-Jen Ko, Akiko Ogawa, Éamonn Ó Tuama, Herbert Schentz, Su Wen, Bert van der Werf .....	156
L. A. Winslow, B. J. Benson, K. E. Chiu, P. C. Hanson, T. K. Kratz .....	166
Jianting Zhang, Kate S. He' and Michael Gertz .....	172
<b>Contributed Poster Abstracts</b> .....	178
Ankit Agarwal, James Beach, Julio Ibarra .....	178
Karen S. Baker, Nicole E. Kaplan, Inigo San Gil, Margaret O’Brien, Florence Millerand ...	178
David Balsiger, Barbara Benson, Jeff Maxted, Luke Winslow .....	179
Chad W Berkley, Shawn Bowers, Matthew B. Jones, Mark Schildhauer .....	180
Shira Bezalel .....	180

Daniel Crawl, Peter Cornillon, Ilkay Altintas, Nathan Potter, James Gallagher, Mark Schildhauer, Matthew B. Jones .....	181
Scott C. Curran, Mike Kearsley.....	182
Judith B. Cushing, Nicole E. Kaplan, Christine Laney, Carri LeRoy, Juli Mallett, Ken Ramsey, Kristin Vanderbilt, Lee Zeman, Judith B Cushing, Juli Mallett, Lee Zeman, Nicole Kaplan, Christine Laney, Ken Ramsey, Kristin Vanderbilt.....	183
Michael Daigle, Matthew B. Jones, Benjamin Leinfelder, Shaun Walbridge, Jing Tao .....	183
Peter C. Griffith, Lisa E. Wilcox, Amy L. Morrell.....	184
Eric S Hersh, David R Maidment .....	184
Vivian Hutchison .....	185
Nicole E. Kaplan, Kristin L. Vanderbilt, Lee Zeman, Judy B. Cushing, Christine Laney, Juli Mallett, Ken Ramsey, Jincheng Gao, Judith Kruger, Carri LeRoy, Daniel Milchunas, Esteban Muldavin .....	185
Mason A. Kortz, James E. Conners, Karen S Baker.....	186
Adam Mellor.....	186
William Michener, Suzie Allard, Paul Allen, Peter Buneman, Randy Butler, John Cobb, Robert Cook, Patricia Cruse, Bruce Dancik, Ewa Deelman, David DeRoure, Mindy Destro, Cliff Duke, Charles Fox, Mike Frame, Stephanie Hampton, Carole Goble, Nancy Grimm, Donald Hobern, Peter Honeyman, Jeffery Horsburgh, Vivian Hutchison, Matthew B. Jones, Steve Kelling, Jeremy Kranowitz, John Kunze, Hilmar Lapp, David Leslie, Jr., Bertram Ludaescher, Thomas Moritz, Lorraine Normore, Robert Peet, Ricardo Pereira, Line Pouchard, Jim Reichman, Hannu Saarenmaa, Robert Sandusky, Ryan Scherle, Mark Schildhauer, Mark Servilla, Kathleen Smith, Carol Tenopir, Paul Uhlir, Dave Vieglais, Todd Vision, Jake Weltzin, Von Welch, Bruce Wilson .....	187
John H. Porter .....	188
Steve Rentmeester.....	188
Aaron Schultz, Matthew B. Jones, Timothy McPhilliips, Sean Riddle, David Welker .....	189
Jing Tao, Matthew B. Jones, David Vieglais, Arcot Rajasekar, Lucas Gilbert, Benjamin Leinfelder .....	189
Theresa Valentine .....	190
Shaun Walbridge, Mark Schildhauer, Jim Regetz, Matthew B. Jones, Rick Reeves .....	190
Lynn R. Yarmey, Karen S. Baker .....	191

## Program Committee

---

### Conference Co-Chairs

Corinna Gries  
Matthew B. Jones

### Program Committee

Todd Ackerman  
Barbara Benson  
James Brunt  
Barrie Collins  
Judy Cushing  
Mike Frame  
Peter Griffith  
Christopher Jones  
Nicole Kaplan  
Raymond A McCord  
Eda Celina Melendez-Colom  
William Michener  
Margaret O'Brien  
Mark Schildhauer  
Wade Sheldon  
Jonathan Walsh  
Bruce Wilson

## Sponsors

---

Long-Term Ecological Research Network (LTER)  
National Center for Ecological Analysis and Synthesis (NCEAS)  
Oak Ridge National Laboratory  
USGS National Biological Information Infrastructure (NBII)  
NASA Terrestrial Ecology Program  
Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO)  
Evergreen State University



## **EIM 2008 Program at a Glance**

### **Wednesday, September 10, 2008**

- 7:00 AM **Registration and Coffee**
- 8:30 - 10:30 AM **Papers: Sensors and Sensor Networks**
- 11:00 AM - 12:00 PM **Keynote Address: James Clark, Duke University**
- 1:15 - 3:00 PM **Papers: Metadata and Data Management Systems**
- 3:30 - 5:30 PM **Poster Session and Reception**
- 7:30 - 9:00 PM **Panel Discussion: Challenges and solutions to managing, accessing, and using sensor data**

### **Thursday, September 11, 2008**

- 8:30 - 10:15 AM **Papers: Data Integration, Analysis and Visualization**
- 10:45 - 11:45 AM **Keynote Address: Tony Beasley, National Ecological Observatory Network (NEON)**
- 1:00 - 2:30 PM **Panel Discussion: Achieving global, cross-institutional interoperability of ecological and environmental data and metadata systems**
- 2:45 - 4:30 PM **Papers: Quality Assurance Systems**

# Environmental Information Management 2008

Sept 10-11, 2008  
University of New Mexico

## Wednesday, September 10, 2008

7:00 AM Registration and Coffee

(Student Union Building: outside Ballroom A)

8:15 AM Welcome to the conference

C. Gries and M. Jones

### Sensors and Sensor Networks

Moderator: M. Jones (Student Union Building: Ballroom A)

8:30 AM Bainbridge, Rehbein, Feather, Eggeling

Sensor Networks on the Great Barrier Reef – managing marine sensor data

8:45 AM Conover, Regner, Maskey, Lu, Li, Goodman

Evolving Sensor Web Protocols for Sensor Data Management

9:00 AM Benson, Winslow, Arzberger, Carey, Fountain, Hanson, Kratz, Tilak

Meeting the challenges of an international, grassroots organization of sites deploying sensor networks: the Global Lake Ecological Observatory Network (GLEON)

9:15 AM Barseghian, Altintas, Jones

Accessing and Using Sensor Data within the Kepler Scientific Workflow System

9:30 AM Liu, Hill, Abdelzاهر, Heo, Choi, Minsker, Fazio

Virtual Sensor-Powered Spatiotemporal Aggregation and Transformation: A Case Study Analyzing Near-Real-Time NEXRAD and Precipitation Gage Data in a Digital Watershed

9:45 AM Vande Castle, Servilla

Streaming Sensor Data from Space: Acquiring and Managing Direct Broadcast Satellite Data for Sites of the Long Term Ecological Research Network

10:00 AM Henshaw, Bierlmaier, Bond, O'Connell

Building a "Cyber Forest" in Complex Terrain at the Andrews Experimental Forest

10:15 AM Questions + Discussion

10:30 - 11:00 AM BREAK

**Keynote Address: James Clark, Duke Univ.**

(Student Union Building: Ballroom A)

11:00 -12:00 PM

Title to be determined

12:00 - 1:15 PM **LUNCH**





# Environmental Information Management 2008

Sept 10-11, 2008  
University of New Mexico

## Wednesday (continued)

### Metadata and Data Management Systems

Moderator: J. Cushing (Student Union Building: Ballroom A)

1:15 PM Leinfelder, Tao, Costa, Jones,  
Servilla, O'Brien, Burt

Using metadata for loading and querying heterogeneous scientific data

1:30 PM Palanisamy, Wilson, Devarakonda,  
Green

Mercury: A Distributed Metadata Management, Data Discovery and Access System

1:45 PM Moore, Walther, Cooper

Data Collaboration for Large-Scale Regional Surveys in Southern California

2:00 PM Rugg

Archival Data Formats - Archivists and Users

2:15 PM Babin, Hu

A Tale of Two Observing Systems: Visualization of Real-time Coastal Ocean Data on the Web

2:30 PM Winslow, Benson, Chiu, Hanson,  
Kratz

Vega: A Flexible Data Model for Environmental Time Series Data

2:45 PM Questions + Discussion

3:00 - 3:30 PM BREAK

### Poster Session

(Student Union Building: Ballroom B)

3:30 - 5:00 PM Poster Session

5:00 - 5:30 PM **Reception** and continued Poster session

5:30 - 7:30 PM **DINNER BREAK**

### Panel Discussion

Moderator: Barbara Benson (Student Union Building: Ballroom A)

7:30 - 9:00 PM

Challenges and solutions to managing, accessing, and using sensor data



# Environmental Information Management 2008

Sept 10-11, 2008  
University of New Mexico

## Thursday, September 11, 2008

8:00 - 8:30 AM Coffee

### Data Integration, Analysis and Visualization

Moderator: Corinna Gries (Student Union Building: Ballroom A)

8:30 AM Zhang, He, Gertz

LEEASP: A Linked Environment of Coordinated Multiple Views for  
Exploratory Analysis of Large-Scale Species Distribution Data

8:45 AM Servilla, Costa, Laney, San Gil,  
Brunt

The EcoTrends Web Portal: An Architecture for Data Discovery and  
Exploration

9:00 AM Porter, Smith

Live from the Field: Managing Live-Image Databases at the Virginia Coast  
Reserve

9:15 AM Cushing, Kaplan, Laney, Mallett,  
Ramsey, Vanderbilt, Zeman, Gao,  
Kruger, Leroy, Milchunas, Muldavin

Integrating Ecological Data: Notes from the Grasslands ANPP Data  
Integration Project

9:30 AM Vanderbilt, Blankman, Guo, He, Li,  
Lin, Lu, Ko, Burke, Ogawa, Ó  
Tuama, Schentz, Su, van der Werf

Building an Information Management System for Global Data Sharing: A  
Strategy for the International Long Term Ecological Research (ILTER)  
Network

9:45 AM Kennedy, Remillard, Henshaw,  
Duncan, Bond

Converting data to information: Coupling lab-level database functionality  
with primary LTER data archiving systems

10:00 AM Questions + Discussion

10:15 - 10:45 AM BREAK

### Keynote Address: Tony Beasley, National Ecological Observatory Network (NEON)

10:45 - 11:45 AM

NEON: Project Status & Technical Developments

11:45 - 1:00 PM LUNCH



# Environmental Information Management 2008

Sept 10-11, 2008  
University of New Mexico

## Thursday (continued)

### Panel Discussion

1:00 - 2:30 PM

Moderator: Mark Schildhauer (Student Union Building: Ballroom A)  
Achieving global, cross-institutional interoperability of ecological and environmental data and metadata systems

2:30 - 2:45 PM BREAK

### Quality Assurance Systems

Moderator: Bruce Wilson (Student Union Building: Ballroom A)  
Challenges of AVHRR Vegetation Data for Real Time Applications  
An Overview of Quality Control Procedures for Buoy Data at the National Oceanic and Atmospheric Administration's (NOAA) National Data Buoy Center (NDBC).  
Dynamic, Rule-based Quality Control Framework for Real-time Sensor Data  
Processing and quality control of kelp forest community survey data

2:45 PM Brown

3:00 PM Sears

3:15 PM Sheldon

3:30 PM O'Brien, Harrer

3:45 PM Langman, Hanson, Carpenter, Chiu, Hu

4:00 PM Hu, Babin

4:15 PM Questions + Discussion

4:30 PM **Concluding remarks**

Jones and Gries



# Environmental Information Management 2008

Sept 10-11, 2008  
University of New Mexico

## List of Posters

- 133 Agarwal, Beach, Ibarra  
Architectural and Functional Requirements for an Environmental Sensor  
Network Computing Platform for Terrestrial Biological Research Stations and  
Ecological Observatories  
LTER Information Managers: A Community of Practice
- 139 Baker, Kaplan, San Gil, O'Brien,  
Florence Millerand  
Recent Information Management System Enhancements at the North  
Temperate Lakes LTER
- 132 Balsiger, Benson, Maxted, Winslow  
Improving metadata search efficiency by enabling semantic queries  
Environmental Data Upload and Visualization Tools
- 124 Berkley, Bowers, Jones, Schildhauer  
140 Bezalel  
An Integrated Framework for Hybrid and Adaptive Modeling of Sea Surface  
Temperature: A Workflow-Based Approach to Comparison
- 136 Crawl, Cormillon, Altintas, Potter,  
Gallagher, Schildhauer, Jones  
144 Curran, Kearsley  
Electronic Collection of Vegetation Mapping Data within the Grand Canyon  
National Park
- 148 Cushing, Kaplan, Laney, LeRoy,  
Mallett, Ramsey, Vanderbilt, Zeman  
149 Cushing, Mallett, Zeman, Kaplan,  
Laney, Ramsey, Vanderbilt  
141 Daigle, Jones, Leinfelder, Walbridge,  
Tao  
Cross-Site Analysis of Abiotic Drivers and ANPP at Five Grasslands Sites
- 71 Griffith, Wilcox, Morrell  
123 Hersh, Maidment  
16 Hutchison  
Problems and Solutions in Species-Coded Data: Best Practices and Common  
Issues  
Simplified deployment of the Metacat data and metadata system
- 128 Kaplan, Vanderbilt, Zeman, Cushing,  
Laney, Mallett, Ramsey, Gao, Kruger,  
LeRoy, Milchunas, Muldavin  
138 Kortz, Connors, Baker  
The North American Carbon Program Google Earth Collection  
Managing Information for Environmental Flows in Texas  
USGS NBII Releases Re-Designed Interface for Metadata Clearinghouse  
A Team Approach to Data Synthesis: The Playbook for Creating a Centralized,  
Dynamic, and Sustainable ANPP Database
- 41 Mellor  
Abstracting Functionality and Access: Facilitating Data System Manageability  
and Site Coordination  
North of Ireland Coastal Monitoring Programme - QA for an operational  
network of moored oceanographic instruments



# Environmental Information Management 2008

Sept 10-11, 2008  
University of New Mexico

## List of Posters (con't)

- 127 Michener, Allard, Allen, Buneman, Butler, Cobb, Cook, Cruse, Dancik, Deelman, DeRoure, Destro, Duke, Fox, Frame, Hampton, Goble, Grimm, Hobern, Honeyman, Horsburgh, Hutchison, Jones, Kelling, Kranowitz, Kunze, Lapp, Leslie, Ludaescher, Moritz, Normore, Peet, Pereira, Pouchard, Reichman, Saarenmaa, Sandusky, Scherle, Schildhauer, Servilla, Smith, Tenopir, Uhler, Vieglais, Vision, Weltzin, Welch, Porter
- 135 Porter
- 143 Rentmeester
- 147 Schultz, Jones, McPhillips, Riddle, Welker
- 130 Tao, Jones, Vieglais, Rajasekar, Gilbert, Leinfelder
- 126 Valentine
- 142 Walbridge, Schildhauer, Regetz, Jones, Reeves
- 137 Yarmey, Baker
- Building the Framework for a Virtual Data Center for Ecology and the Environmental Sciences
- Implementing an Automated Processing System for Low-Frequency Streaming Data Using an Eclectic Approach
- A Framework for Defining and Enforcing Multiple Validation Environments (i.e. Protocols) within Aquatic Ecology
- Promoting Community Contributions with Highly Configurable Component Based Software, A Kepler Architecture
- EarthGrid Web Services for Accessing Heterogeneous Data Systems
- Integrating Google Earth and Internet Mapping into Your Website
- Web-based collaboration in an ecology think-tank
- Information Infrastructure: Emergent Roles, Responsibilities and Practices



## **Contributed Oral Papers**

---

## **A TALE OF TWO OBSERVING SYSTEMS: VISUALIZATION OF REAL-TIME COASTAL OCEAN DATA ON THE WEB**

**Brenda Leroux Babin<sup>1</sup> and Lei Hu<sup>2</sup>**

<sup>1</sup>Louisiana Universities Marine Consortium, Chauvin, LA, 985-851-2878, [bbabin@lsu.edu](mailto:bbabin@lsu.edu)

<sup>2</sup>Dauphin Island Sea Lab, Dauphin Islands, AL, 251-861-7533, [luh@disl.org](mailto:luh@disl.org)

### **Abstract**

Many coastal environmental monitoring systems exist around the United States. With the advent of the Integrated Ocean Observing System (IOOS) initiatives it is important that these systems find ways to maintain the data locally while providing a distributive approach to disseminating and visualizing the data. Louisiana Universities Marine Consortium and Dauphin Island Sea Lab, two Gulf coast marine labs, were establishing coastal environmental monitoring systems. In 2000, the data managers from the two labs came together and decided to create two systems mirrored after each other, so that the data would be stored locally at each lab and could be readily accessed at the other. Microsoft SQL Server, active server pages, and the web served as the backbone for this project. The two labs implemented ChartDirector for visualization. This backbone allows users on the web to visualize the data from the two labs side by side on the web. The system is also designed in such a way to facilitate the implementation of open source tools for participation in IOOS.

**Keywords:** environmental monitoring, data visualization, ocean observing systems

### **1. Introduction**

The Integrated Ocean Observing System (IOOS) is a system of smaller observing systems that provides continuous quality controlled real-time data. Ocean.us, the national office for the IOOS, identifies seven goals for the IOOS including improving predictions of climate change and weather and their effects on coastal communities and the nation; improving the safety and efficiency of maritime operations; mitigating the effects of natural hazards more effectively; improving national and homeland security; reducing public health risks; protecting and restoring healthy coastal ecosystems more effectively; and enabling the sustained use of ocean and coastal resources (Malone 2000). The success of IOOS requires that individual systems provide a standardized method of data delivery while maintaining autonomy of data maintenance within the local system.

Louisiana Universities Marine Consortium (LUMCON) and Dauphin Island Sea Lab (DISL) maintain two such systems which provide meteorological and water quality parameters along the Louisiana and Alabama coasts respectively. Some of these parameters include air temperature, wind speed, wind direction, water temperature, chlorophyll, turbidity, and salinity. Users of these systems included marine scientists and local managers requiring that the data are available in real-time on the World Wide Web and that the archive data be easily accessible locally. Focusing on the reasons for collecting data is the first step to ensuring the users' needs are met.

Our challenge was to design a system to fit the needs of our data users while maintaining the ability to adapt our systems to the IOOS standards which were still being developed. From the sensor to the web, this goal remained our focus. In 2000, we decided to mirror the two

marine labs after one another and create a “mini” coastal-ocean observing system. In order to achieve this goal the background and infrastructure had three main requirements. First, the data had to be collected and stored locally. The data had to be available on the web within one minute of collection. And, the data visualization had to allow for side by side displays of data from both systems.

## 2. Methods

### 2.1 Study Area

LUMCON’s Environmental Monitoring System collects and archives real-time meteorological and hydrographic data to provide a broad community of scientists, educators, students, and the public with quality-controlled environmental data from Louisiana’s Gulf Coast. The Louisiana Universities Marine Consortium (LUMCON) was formed in 1979 to coordinate and stimulate Louisiana's activities in marine research and education. LUMCON provides coastal laboratory facilities to Louisiana universities, and conducts research and educational programs in the marine sciences. LUMCON established a coastal environmental monitoring system to bring coastal information to scientists, educators, students, and the public throughout the state. All of the data are freely available in real-time via the Internet. Six remote monitoring stations, located along the southeastern Louisiana coast (Fig. 1), collect environmental data from an array of meteorological and hydrographical instruments.

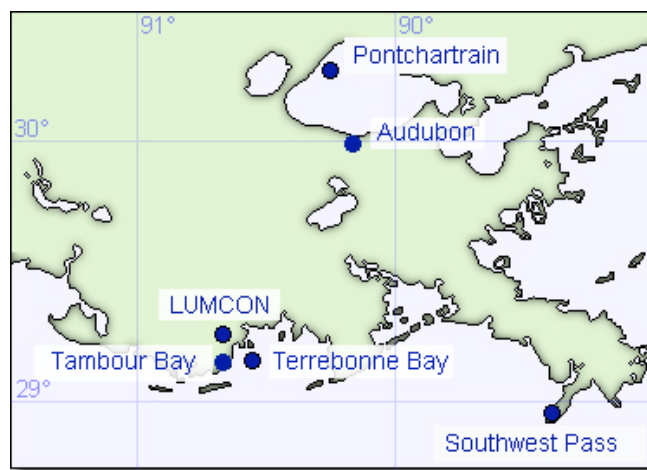


Figure 1. LUMCON’s environmental monitoring stations located throughout southeast Louisiana.

Dauphin Island Sea Lab (DISL), founded in 1971 by the Alabama State Legislature, is Alabama’s marine science education and research laboratory. Located on Dauphin Island, a barrier island in the Gulf of Mexico, the DISL primarily serves the twenty-one four-year colleges and universities of Alabama through its college summer courses and graduate programs. It also offers Discover Hall Programs, which encompasses K-12 field programs, teacher-training, and public outreach. The research programs of the DISL range from biogeochemistry to oceanography to paleoecology. Starting in 2000, the DISL collaborated with LUMCOM to create the environmental monitoring system in Mobile Bay. The system includes three stations, Dauphin Island, Meaher Park, and Middle Bay Light (Fig. 2). DISL also collects the data for two National Estuarine Research Reserve sites Weeks Bay and Wolf Bay. These data are also included as part of the DISL system.

### 2.2 Data Collection

At each station, Campbell Scientific micro-loggers collect data from the instruments. Once per minute the data are averaged and transmitted to the marine labs via spread spectrum radio transmission, cellular modem and/or direct Internet connection. The meteorological parameters include wind speed and direction, air temperature, relative humidity, precipitation,



barometric pressure, solar radiation, and quantum radiation. The hydrographic parameters include water temperature, water height, salinity, conductivity, turbidity, dissolved oxygen, and chlorophyll concentration.

### 2.3 Data Storage

At the both marine labs, the data are then stored in a Microsoft SQL server database as well as archived in comma separated value (CSV) files. Although not a requirement for this project, the schema of these two SQL servers are very similar. Current values are displayed in real-time on the website. One-day, five-day, thirty-day and yearly graphs are generated with current information when selected by a web visitor. A user can also download archived data by using a form to subset the data according to a set of criteria from the SQL server or by accessing the CSV raw text files directly through a web browser or ftp. At both marine labs, an OpenDap Server provides the data in NetCDF format. Every hour, the servers at LUMCON automatically send the data to the National Data Buoy Center (NDBC) via ftp, and the data become part of the National Weather Service (NWS) data stream. In 2005 the DISL started sending data in the format of XML every 30 minutes to the National Data Buoy Center (NDBC).



Figure 2. Dauphin Island Sea Lab’s environmental monitoring stations located on the Alabama coast.

### 2.4 Data Display

The code to generate the graphs for both systems leverages Advanced Software Engineering Limited's ChartDirector. ChartDirector is a commercial application providing professional chart component for windows and web applications. Although there are several programming language editions of ChartDirector, the active server page (ASP) edition allows the programmer to create chart objects and display these as standard graphics on a web page. The active server page program graph.asp written in Microsoft Visual Basic script generates the charts based on parameters passed to the program in the URL and sends the chart in jpg format to the browser.

By simply using the <img> tag in html and pointing to chart.asp as the “src” parameter anyone can leverage these charts in web pages using simple html coding ().

<b>Table 1.</b> Parameters passed to the program.		
Parameter	Description	Possible Values
param	The name of the parameter to chart	Airtemp, precip, pressure, relhumid, solarrad, quantumrad, winddir10m, windspeed10m, watertemp, salinity, turbidity, waterht, waveht, flouro, DO
stationID	A unique three digit identifier	101,102,etc.

	for each station	
ChartYear	Year of the data to include in the chart	2000-2008
jday	Julian day of last day in the chart	1-365
ChartType	number of days to include in the chart	1-365
SciUnits	The type of units to use	1=Scientific Units, 2=English Units

### 3. Results

The technique of using the URL in the image tag allows for real-time display of data from the two systems on the same web page. Thus by varying the parameters two stations can be displayed side by side (Fig. 3). This system allows for the display of these data on other websites in real-time.

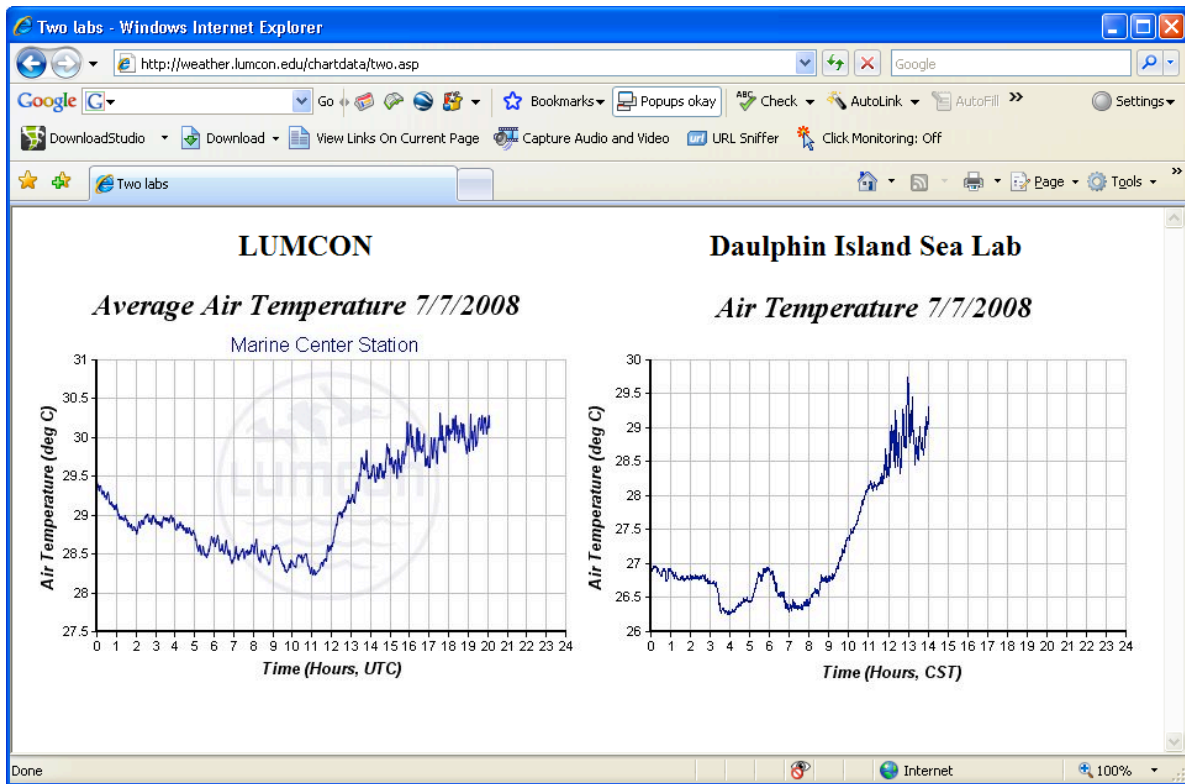


Figure 3. Screen capture of side by side display of real-time graphs from a LUMCON station and a DISL station. <http://weather.lumcon.edu/chartdata/two.asp>

Another advantage of our system is the ability to adapt and add components. We recently upgraded all of our water quality sensor technology from Yellow Springs Inc. (YSI) 6600 extended deployment sondes to the 6600 V2. This upgrade included the new YSI optical dissolved oxygen sensor. The addition of dissolved oxygen to our sensor array created the necessity of adding dissolved oxygen to our visualization. This project was completed by simply adding a record for dissolved oxygen in the "ParameterInfo" table in the database in SQL server creating all of the necessary parameter values to now add dissolved oxygen to the system.

#### **4. Discussion**

We can now display graphs from two different stations from different systems on the same webpage. The prototype webpage two.asp shows the current temperature graph from DISL's Dauphin Island site and the current temperature graph from LUMCON's Marine Center site side by side. These graphs are created on-the-fly when the user accesses the webpage keeping the data display current.

One of the challenges DISL ran into was how to use this program for the profiling sensor in Mobile Bay. This water quality sensor travels vertically in the water column each hour taking a measurement and set depth intervals. Unfortunately, a separate graphing program had to be created for this particular site and some standardization was lost.

After other environmental monitoring systems have adapted our protocol for displaying data in real-time online, we have been able to incorporate this data into our websites. One example of such a system is WAVCIS ([wavcis.lsu.edu](http://wavcis.lsu.edu)). The WAVCIS program created their own data display system using Chardirector; however, it uses the same basic principles that we are using in our system and the graphs can be displayed in other web pages. This is how [www.gulfhypoxia.net](http://www.gulfhypoxia.net) is displaying dissolved oxygen data from two sites in the northern Gulf of Mexico. Using a specifically designed URL, the gulfhypoxia site accesses the WAVCIS graphs and displays them on-the-fly when the user accesses the page.

One of the next steps is to "map" data to allow for spatial visualization of the data from these two systems. We are working on a plan to use the NDBC-XML format to allow for on-the-fly data aggregation into a "google" type mapping system.

As IOOS standards developed along the lines of open source standards, both systems have been able to adapt to these standards. IOOS data management and communication standards have moved from OpenDAP serving NETCDF standard data to Open Geospatial Consortium (OGC) standards. The design of both the LUMCON and the Dauphin Island systems have been enabled them to implement the Open DAP servers and to continue to adapt as the standards continue to change. This visualization project is just one example of the adaptability of these two systems illustrating how these systems can maintain their autonomy and still be part of an integrated ocean observing system.

#### **Acknowledgements**

Equipment for construction and instrumentation for LUMCON's Environmental Monitoring System was obtained through grants from a variety of sources including: NOAA, EPA, NASA, USGS and the State of Louisiana. Maintenance and operation costs have been borne by various grants and LUMCON.

Funding from the Mobile Bay National Estuary Program, EPA's Gulf of Mexico Program, Alabama Department of Conservation, State Land Division, the University of South Alabama and the Coastal Impact Assistance Program was used to establish and maintain the Dauphin Island Sea Lab System.

#### **References**

Malone, T.C. and Cole, M., 2000. Toward a global scale coastal ocean observing system. *Oceanography*. 13: 7-11.

## **SENSOR NETWORKS ON THE GREAT BARRIER REEF – MANAGING MARINE SENSOR DATA**

**Scott Bainbridge, Mark Rehbein, Gavin Feather and Damien Eggeling**

Australian Institute of Marine Science, PMB 3 MC, Townsville Qld 4810 Australia

### **Abstract**

The Great Barrier Reef Ocean Observing System project is deploying sensor networks at seven sites along the Great Barrier Reef in north-eastern Australia. The project has a strong data focus and is actively developing systems to manage the data collected. A data schema based on deployments has been developed with a deployment hierarchy of platforms (e.g. buoys, moorings), instruments such as loggers and the sensors themselves. Supporting schema entities include an equipment register that holds the details of the equipment deployed, service history and calibration entities and a sensor relationship entity that holds details of how individual sensors relate to each other. Metadata is collected using the Australian Marine Community Profile of ISO-19115 using the GeoNetwork open-source software. A framework for quality control has been developed that at the lowest level uses the International Oceanographic Commission (IOC) / International Oceanographic Data and Information Exchange (IODE) set of quality control flags and a simple rules based system to deliver a 'Level-0' quality controlled product. Higher levels are also defined where manual corrections and complex processing are applied to create higher level data products or versions of the data. The Open GIS Consortium Sensor Web Enablement framework has been chosen for data exchange and representation. SensorML is used to describe the sensor systems while Observation and Measurement ML is being investigated for the data itself. The usability of these XML standards is an issue as many are complex and the supporting tools and software are still under development. The data will be made available directly as web services either as spatial web services or as a pure data stream service.

**Keywords:** Great Barrier Reef, Sensor Networks, Data Management, Real-Time Data, Marine Data

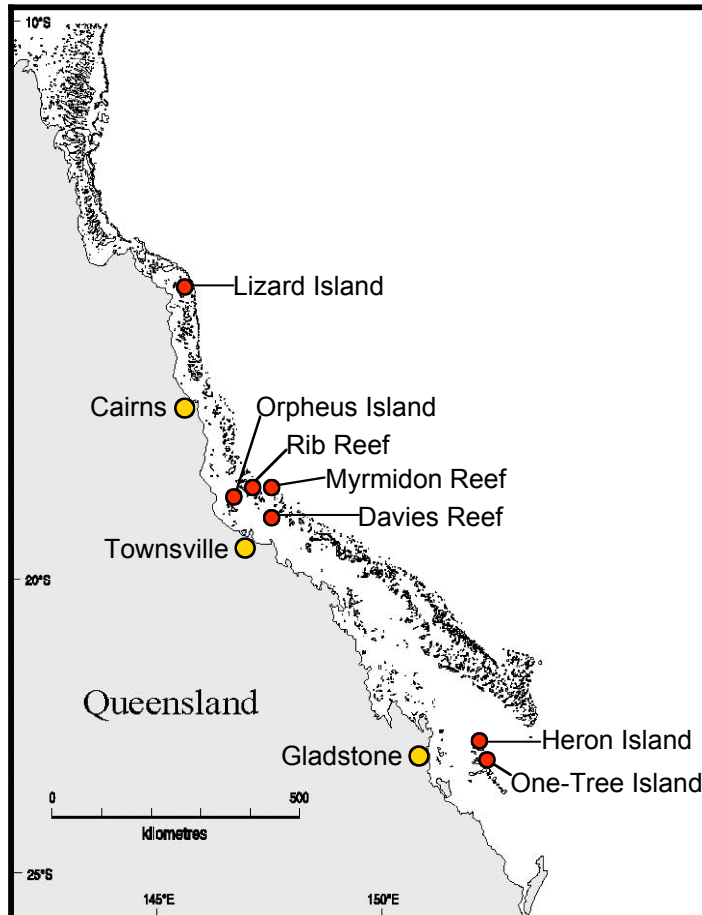
### **1. Introduction**

The Great Barrier Reef Ocean Observing System (GBROOS) seeks to measure and monitor the impact of water flowing from the Coral Sea onto the Great Barrier Reef (GBR) in north-eastern Australia and then to track this water as it flows through the reef matrix and forms the major currents that run south as the East Australian Current and north as the Hiri Current. GBROOS is a geographic node under the Australian Integrated Marine Observing System (IMOS) project (IMOS, 2008).

One component of GBROOS is the deployment of reef based sensor networks at seven sites on the GBR, four of which are sites with an island research station associated with the reef (Heron and One Tree Islands in the southern GBR, Orpheus Island in the central GBR and Lizard Island in the northern GBR) and three are isolated reefs (Rib, Davies and Myrmidon Reefs in the central GBR). Figure One shows the location of the sites.

At each reef an above water wireless network is created into which a number of sensor buoys are deployed, the data flows back via the wireless network to a base station located on the reef and then, via broad-band IP links, to the mainland where the data is stored and managed.

Figure 1. Location of the sensor network sites.



## 2. Methods and Techniques

At seven sites along the Great Barrier Reef sensor networks will be deployed. Each deployment consists of a broad-band IP link back to the mainland, a base station and an on-water wireless network. The base station acts to interface the back-haul IP based network and the on-reef pack-bus network.

The on-reef network is created using spread-spectrum radios mounted on a series of poles located in the reef lagoon every two kilometres.

Sensor buoys are then deployed into this network using spread-spectrum radios to talk back to the base station. The floats have a Campbell Scientific CR1000 logger into which a range of sensors are attached using RS-232, SDI-12, one-wire or inductive modem interfaces.

The Campbell Scientific LoggerNet software is used to communicate to the loggers and sensors and to monitor the system. The

LoggerNet software creates a set of data files that are read by an in-house Java program which inserts the data into an Oracle database. The Java program is a simple harvester and uses static configuration information to link the data files from LoggerNet to the Sensor and Deployment Id's in the data schema.

A range of in-house Java programs are then used to manipulate the Oracle data such as conducting the quality control processes, monitoring of the system and serving the data. A separate stream of non-controlled data is split off to a public server running the DataTurbine (DataTurbine, 2008) product for people interested in real-time streaming data.

## 3. Results and Discussion

Sensor networks promised large amounts of relatively cheap data and many data management strategies proposed reflect this. The engineering and logistics issues that the marine environment imposes mean that marine sensors will never be 'cheap and cheerful'. For the GBROOS project, each sensor (such as a fifty cent off the shelf thermistor) will cost around US\$15,000 - \$30 000 to deploy and service over a life of three years.

Along with the equipment cost there is the opportunity cost. When the network is fully deployed, the Great Barrier Reef, which covers an area of 344,400 square kilometres (DEH, 2006), will have less than a thousand sensors to measure and monitor a system that is four dimensional, complex at all scales, and which is undergoing complex changes in response to climate and other changes.

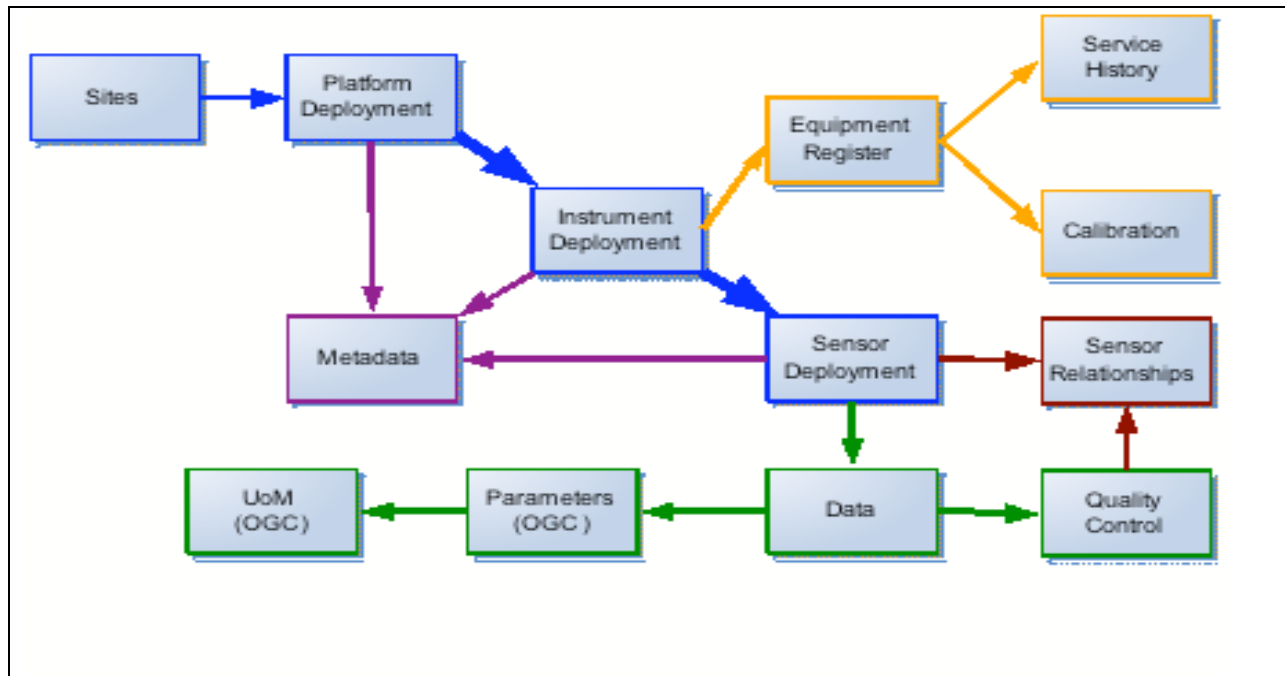
With so few sensors measuring a large dynamic complex system the impact of unreliable or failed equipment is enormous. The science demands high value data and so the systems, sensors and the data management architecture need to reflect the high cost of the data and the high opportunity cost in lost or invalid data.

The approach taken uses relational databases with traditional database schemas and database programming languages. The quality control framework developed looks to conserve as much data as possible using a hierarchy of ‘flags’ to identify data that may be bad or unusable rather than discarding or deleting data. The data management systems need to preserve data where possible and so the systems are design to be data conservative.

### 3.1 Data Schemas

A Data Schema for the project has been developed based on the concept of deployments (Figure Two). A physical structure such as a buoy or mooring is deployed in the water. This is called a platform. On this platform are deployed instruments such as loggers. On the instrument a range of sensors are deployed, these are the devices that measure the real-world phenomena. This schema therefore gives a hierarchy of platforms, instruments and sensors.

Figure 2. Data Schema used for the GBROOS sensor data.



The schema includes the idea of a site at which platforms are deployed. For marine sensors this is somewhat redundant as many deployments are away from fixed locations but as sites are fundamental to terrestrial and coastal work they have been retained.

An equipment register is included where details of all equipment are stored along with their calibration and service histories. All information that is not related to a particular deployment of the equipment is kept in the equipment register; information related to each deployment is kept in the deployment objects.

The data entity is arranged as a UTC date/time value, the Id of the sensor, the resulting measured value, the parameter that the value represents and the units of measure used. This structure reflects that a measurement can be in more than one unit (e.g. temperature in degrees Celsius or degrees Kelvin) and that the same sensor can be used to measure a range of phenomenon (e.g. water temperature, air temperature). Standard values (such as Open GIS Consortium terms) are used for Parameters and Units of Measure (UoM).

A sensor relationship entity is included which serves a number of purposes. Where a dataset is made up of a series of shorter deployments the sensor relationship entity links these together (using 'comes before' or 'comes after' relationships). The second purpose is for quality control. Often an expensive reference sensor is deployed to provide real-time corrections for a series of cheaper sensors; the sensor relationship entity holds this relationship and allows the automated correction of the cheaper sensor data against the reference sensor.

### **3.2 Metadata**

The project has adopted the ISO-19115 metadata standard (ISO, 2003) for spatial data and in particular the Australian Marine Community Profile (MCP) of ISO-19115 (AODC, 2007). The open-source GeoNetwork software (FAO, 2008) is used to enter and edit metadata. The ISO concept of parent and child metadata records is used to deal with the issue of what level or granularity of metadata should be recorded.

Each platform deployment is described by a single metadata record. A child record is then created for each instrument deployed on that platform; a child record of the Instrument record (or grand-child of the platform record) is created for each sensor deployed on the Instrument. The metadata therefore reflects the hierarchy of deployments described by the data schema.

A deployment is defined as an uninterrupted period of operation when consistent data can be expected. If a sensor is removed, replaced, cleaned or altered then that becomes a new deployment even if the platform or instrument is not touched. If a platform is moved or changed then that will trigger a new deployment for all the instruments and sensors on that platform. Each new deployment generates a new metadata record to describe the deployment.

### **3.4 Quality Control**

The project has developed a complex framework for quality control of the data but is struggling with implementing this. The overall philosophy is to keep as much data as possible, to use flags to identify the quality of the data and to give the user enough information so they can decide what data they should use based their individual application.

The framework is based on a series of levels. Level-0 leaves the data intact and implements a simple set of flags based on the UNESCO/IOC/IODE quality control flags (see Table One). A series of simple range checks and other automated tests are done to identify data that is missing (no value collected), wrong (values are illegal) or illogical (values do not make sense). Importantly Level-0 does not alter the data; it just flags it as good or otherwise.

Table 1. Quality Control Flags (UNESCO/IOC/IODE & MAST, 1993).

Flag Value	Meaning
0	No quality control has been applied to the data value
1	The data value appears to be correct
2	The data value appears to be probably good
3	The data value appears to be probably bad
4	The data value appears erroneous
5	The data value has been changed
6 to 8	Reserved for future use
9	The data value is missing

Level-1 and above do involve altering the data such as interpolating missing values, de-spiking, de-trending and so on. As such these represent new versions of the data or new data products. The quality control flags now indicate how the value was obtained; if it is an untouched or raw value or a processed value (such as an interpolated or smoothed value). These higher level flags have yet to be developed. The QC data is stored in the database as a series of pass/fails. Note that the QC data can be large, often many times that of the data.

The intent is to create a rules based system that applies a set of tests and allocates a score based on the outcome. The scores are summed to give a value that sets the IOC/IODE flag. The rules would include simple range checks but also comparisons to other sensors via the Sensor Relationship object, the Service History object to see if servicing or other events had taken place and other checks such as rates of change and historical comparisons. The Kepler software (Kepler Project, 2008) is being trialled to implement the proposed system.

The Data Object holds values for the Level-0 data as the data value and the QC flag value and also holds the same information for Level-1 and Level-2 data and flags. These levels have not been fully defined but it is anticipated that most people would access Level-1 data which has simple de-spiking, interpolation of missing data and inter-sensor corrections applied. Level-0 or raw data would only be used by experienced investigators. Level 2 and higher would be processed data summarised and corrected for long-term trend analysis or other applications.

A final part of the QC system is error. For each point the error for that value is calculated, normally this is the accuracy of the sensor but there maybe times when this value can indicate other sources of error such as when measurements are at the limit of the range of the sensor or when values are interpolated. The error value is important when using differing types of sensors (such as different brands of thermistors); being able to overlay the error on a graph highlights real changes versus those that can be attributed to the sensors themselves.

### 3.5 Data Exchange and Access

Data exchange and access are critical components of the project but this is an area where few mature standards exist. The OCG Sensor Web Enablement (SWE) framework (OGC, 2008) seems to be the most advanced in dealing with sensor data and the project is looking to implement these as they become available. Currently SensorML is used to describe the sensor systems themselves with ISO-19139 used for the metadata.

The project is looking to use web services as the primary delivery mechanism. Currently data is available via REST compliant web services but spatial web services such as Web Feature Services (WFS) and Web Coverage Services (WCS) are being developed. Software such as the



Deegree project (Lat/Long GmbH, 2008) are being trialled to expose the database directly to WFS/WCS compliant clients without the need for a spatial or map server.

For the data itself the project is investigating Observation and Measurement ML but this is still an emerging standard. One issue with many of these data schemas is that they put the data in a comma separated block wrapped by XML. The need to associate point level information, such as quality control and error information, means that the schema needs to support full XML down to the level of each data point.

### **3.6 Issues and Opportunities**

The work so far has raised a number of issues. The main one is a lack of standards and tools to implement the standards that do exist. Many of the standards are complex and not targeted at the simple needs of most users. There is also a feeling that we are 're-inventing the wheel' and that many of the issues we have identified have most likely been (or should have been) solved by others.

The particular issues identified include:

- Lack of standards for quality control including an agreed to approach and framework (such as agreed to flags, processing levels, schemas, tools and so on);
- Issues of how to deal with high-volume / high-value data and the scalability (or lack of scalability) of traditional database based processing systems and architectures;
- What to do with data such as video and images, how to integrate these into simple scalar data;
- Getting the tools in place to support standards such as the OCG SWE standards, the ISO-19115/19139 metadata standard, SensorML and spatial web data services.

With these issues come opportunities. The marine community is small and has shown an ability to work together; similarly the informatics community has a long record of successful collaborative work. The challenge is to identify common issues and to develop a community that can develop the frameworks and standards to deliver standards-compliant data services and information products.

## **4. Conclusion**

This paper presents what the GBROOS project hopes to achieve and the issues and opportunities that we have identified. The GBROOS project will have substantial deployments by the end of 2008 and will need to manage data from a range of sites and sensors. The opportunity is there to work with the international community on the standards, processes and tools to deliver open standards-compliant data systems and information products.

## **Acknowledgements**

The GBROOS project is funded under the Australian Integrated Marine Observing System (IMOS) with the sensor network component funded by the Australian Federal Government, the Queensland State Government along with a number of other research partners. The project is managed by the Australian Institute of Marine Science (AIMS).

Work on the Island Research Stations is supported by the Tropical Marine Network which includes the University of Queensland (Heron Island), the University of Sydney (One Tree Island), James Cook University (Orpheus Island) and the Australian Museum (Lizard Island).

### References

- AODC, 2007. <http://www.aodc.gov.au/files/MarineCommunityProfilev1.3.pdf>
- DataTurbine, 2008. <http://www.dataturbine.org/>
- DEH – Australian Department of Environment and Heritage, 2006. Review of the *Great Barrier Reef Marine Park Act 1975*. <http://www.environment.gov.au/coasts/publications/gbr-marine-park-act.html>
- FAO, 2008. <http://geonetwork-opensource.org/>
- IMOS, 2008. <http://www.imos.org.au>
- ISO, 2003. Geographic Information – metadata.  
[http://www.iso.org/iso/iso\\_catalogue/catalogue\\_tc/catalogue\\_detail.htm?csnumber=26020](http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=26020)
- Kepler Project, 2008. <http://kepler-project.org/Wiki.jsp?page=KeplerProject>
- Lat/Long GmbH, 2008. <http://www.deegree.org/>
- OGC, 2008. Sensor Web Enablement Working Group.  
<http://www.opengeospatial.org/projects/groups/sensorweb>
- UNESCO/IOC/IODE & MAST, 1993. Manual of Quality Control Procedures for validation of Oceanographic Data - UNESCO Manuals and Guides 26

## ACCESSING AND USING SENSOR DATA WITHIN THE KEPLER SCIENTIFIC WORKFLOW SYSTEM

Derik Barseghian<sup>1</sup>, Ilkay Altintas<sup>2</sup>, Matthew B. Jones<sup>1</sup>

<sup>1</sup> National Center for Ecological Analysis and Synthesis, University of California Santa Barbara <sup>2</sup> San Diego Supercomputer Center, University of California San Diego {barseghian, jones}@nceas.ucsb.edu, [altintas@sdsc.edu](mailto:altintas@sdsc.edu)

### Abstract

Environmental sensor networks are now commonly being deployed within environmental observatories and as a component of numerous smaller-scale ecological and environmental experiments. Effectively using streaming data from these sensor networks in near real-time proposes a set of technical challenges that is difficult for scientists to overcome and severely limits the adoption of automated sensing technologies in environmental science. The Realtime Environment for Analytical Processing (REAP) project addresses these technical challenges related to accessing and using heterogeneous streaming sensor data from within the Kepler scientific workflow system. Scientific workflow systems can be used to access, stream, and analyze data from observatory networks and archives. Integrated access to both near real-time data streams and data archives from within the Kepler scientific workflow system facilitates sophisticated analysis and modeling with these data sources. We outline applications of sensor-enabled scientific workflows for a terrestrial ecology use case.

**Keywords:** Scientific Workflows, Sensors, Near Real-Time Data Access, Data Analysis, Terrestrial Ecology

### 1. Introduction and Motivation

Scientific workflows are representations of generally one, but sometimes more, process(es) in the scientific method. They combine data and computational procedures into a configurable, structured set of steps that implement semi-automated computational solutions of a scientific problem. A scientific problem, depending on its focus, could involve ad-hoc processes that a scientist may use to get from raw data to publishable results. Today, scientific workflows are widely being adopted by the scientific and engineering communities due to the advantages they provide on top of existing scripting and visual programming tools. An example advantage of some scientific workflow systems is provenance tracking—the recording of information about workflow execution, intermediate and final results, and the evolution of workflows in time.

Workflows in the Kepler scientific workflow system are mainly composed of sets of *Actors* and *Directors*. In Kepler, workflow authors use local or online Kepler user interfaces to implement an analytical procedure by connecting together a series of workflow components, i.e. “*Actors*”, through which data are processed and flow. Parts of actors that receive *Tokens*, which encapsulate single or multiple data or messages, are called *Ports*. Kepler streamlines the workflow creation and execution process so that scientists can design, execute, monitor, re-run, and communicate analytical procedures repeatedly with minimal effort. Using Kepler, scientists can capture workflows in a format that can easily be exchanged, archived, versioned, and executed (Altintas et al. 2004a).

Scientific workflow systems have been used for accessing data from a variety of sources, including database systems (Altintas et al. 2004a), Grid systems (Ludäscher et al. 2006, Altintas et al. 2003, Altintas et al. 2005), and Web Services (Altintas et al. 2004b). In addition, the Kepler workflow system has built-in tools for accessing heterogeneous environmental data by using details about data content and structure from metadata descriptions that are available in the Knowledge Network for Biocomplexity (KNB), a large-scale, distributed data system. These technologies along with several others in Kepler have been used to solve many scientific problems that require access to data in existing archives. However, one faces new challenges when building analysis workflows using near real-time streaming data. Some of these challenges are accessing and representing nodes in a sensor network as dynamic data sources, synchronization of data coming from different sources, monitoring the health of a network to determine the quality of streaming data, and using streaming data in visualization and analysis applications on-the-fly. Each of these technical challenges in turn brings a number of research and development problems. For example, streaming data from sources with different clocks may need to be synchronized by time before analysis can occur.

The Realtime Environment for Analytical Processing (REAP) project addresses these technical challenges related to accessing and using heterogeneous streaming sensor data from within the Kepler scientific workflow system. In this paper we describe extensions to Kepler that allow users to easily access and utilize streaming data from sensor networks and archived data from the KNB and other data networks.

## **2. Scientific Use Cases**

Our initial development efforts have been in large part driven by the needs of two very different scientific use cases: a terrestrial ecology use case in which near real-time data from terrestrial micrometeorological sensors will aid in a study of plant host populations and their susceptibility to viral pathogens, and an oceanography use case that will compare and match-up remotely sensed sea surface temperature data. As we address the specific needs of these use cases, we do so in a way that promotes the re-use and extension of our work. In this paper we present the requirements of the terrestrial ecology study, and our associated engineering and development efforts.

### **2.1. Terrestrial Ecology**

Non-native annual grasses currently dominate the west coast of the United States in areas historically dominated by perennial native bunchgrasses (Baker 1978, Jackson 1985). The terrestrial ecology use case focuses on the hypothesis that this widespread invasion and sustained domination by non-native annual grasses in California is mediated through interactions with a suite of viral pathogens, the barley and cereal yellow dwarf virus group (B/CYDV), that infects both annual and perennial grasses and is carried by several common aphid species (Halbert and Voegtlin 1995). Although mathematical models and field observations are consistent with this hypothesis (Borer et al. 2007), a thorough test of this hypothesis requires a detailed understanding of grass community phenology, which can be measured with a sensor network that accurately measures ambient meteorological conditions, soil moisture, and biomass accumulation in the grass canopy. This case study requires development of easy-to-use analytical software that supports the analysis and modeling of sensor network data in near real-time to detect local thresholds (e.g. hours exceeding developmental thresholds for aphids), long-term trends (e.g. within- and among-season soil moisture trends), and significant events (e.g.

timing of peak plant biomass). In addition, hypothesis testing for this use case requires integration of sensor data with archived data to assess the relative impacts of disease, plant composition, rainfall, temperature, and soil nutrients on competitive interactions among grasses.

For this study we have deployed hardware that is commonly used by the ecological community in order to develop software against a realistic set of sensor equipment. A Campbell Scientific weather station was deployed at the Baskett Slough National Wildlife Refuge in Dallas, Oregon. The weather station includes a data-logger, a 900mhz spread spectrum radio, and a battery power supply within a weatherproof enclosure. The enclosure, a directional antenna, and a solar panel that serves as power source are mounted on a six-foot tripod (Figure 1). Eight sensors are attached to the data-logger and are mounted on the tripod and ground nearby. A program written in Campbell Scientific's CRBasic language runs on the data-logger, sampling data from the sensors at regular intervals, and a computer at a nearby U.S. Fish and Wildlife Service building periodically establishes radio communication to the weather station and downloads the newly collected data.



Figure 1. REAP weather station. Pictured from left to right, starting at top: anemometer, lightning rod, quantum point sensor, directional antenna, relative humidity and temperature sensor within gill radiation shield, enclosure and solar panel.

### 3. Related Technologies

#### 3.1. OGC Sensor Web Enablement

The Open Geospatial Consortium (OGC) has an initiative called Sensor Web Enablement (SWE) that is "focused on developing standards to enable the discovery, exchange, and processing of sensor observations, as well as the tasking of sensor systems" (Botts et al. 2007). OGC defines the Sensor Web as "web accessible sensor networks and archived sensor data that can be discovered and accessed using standard protocols and application program interfaces (APIs)" (Botts et al. 2007). The SWE initiative has established a number of pending OpenGIS Specifications, including Observations & Measurements Schema (O&M), Sensor Model Language (SensorML), Transducer Markup Language (TransducerML or TML), Sensor Observations Service (SOS), Sensor Planning Service (SPS), Sensor Alert Service (SAS), and Web Notification Services (WNS) (Botts et al. 2007).

REAP is a complementary effort to the SWE initiative; while SWE is focused on the development of standards, REAP is focused on providing scientists, network engineers and the public the ability to access and interact with data and services described by these emerging standards from within a scientific workflow environment.

REAP will also provide new or improved interfaces to data and data-streams already available through other technologies such as DataTurbine (Tilak et al. 2007) and Metacat (Jones, 2001).

### 3.2. DataTurbine

In our terrestrial ecology use case we push our sensor measurements into a DataTurbine server. DataTurbine is an open-source data streaming middleware that provides a robust and generic interface for accessing real-time and user-selected time-ranges of data from a diverse set of sensors. DataTurbine also provides server mirroring capabilities and the ability to link servers together in parent-child relationships (Tilak et al. 2007). We plan a another deployment where a DataTurbine will operate in the field, and in this case we will provide reliable server access by mirroring to a more stable location (i.e. to a server running inside a building, directly connected to the Internet). Once we have deployed more weather stations for our ecological study, we will provide a hierarchical view of all data-streams from station-specific DataTurbines through one parent DataTurbine.

In DataTurbine terminology, data providers are called *Sources*, and data consumers *Sinks*. The DataTurbine API provides a means for developing sink and source applications to easily push and pull numeric, binary and image data (Tilak et al. 2007). After purchasing our weather station we developed a small source application to parse and push the numeric sensor data into a DataTurbine server. In DataTurbine, data and their associated time-stamps are stored together in a "channel", each channel accessible independently of others. The data collected from our weather station forms 13 channels, for example an air temperature channel and two volumetric water content channels. After being pushed into our publicly accessible DataTurbine, anyone running a sink client may access the data.

### 3.3. KNB Metacat

Metacat, short for "Metadata Catalog", is a "network-enabled database framework that lets users store, query, and retrieve XML documents with arbitrary schemas in SQL-compliant relational database systems" (Jones, 2001). Many datasets are stored in Metacats around the world and are accessible within Kepler. A workflow author can use the Kepler search panel to find and use data of interest.

Many Metacat datasets consist of ecological data described in Ecological Metadata Language (EML) (Fegraus et al., 2005). Workflows that use EML data from remote Metacat servers, and EML data from local files (using the Kepler "EML 2 Dataset" actor) in conjunction with near real-time sensor data from a DataTurbine server are planned.

## 4. Terrestrial ecology workflows in Kepler

For the terrestrial ecology use case, we have developed three types of workflows that are critical to the scientific study. The first are event detectors, analyzing incoming streaming sensor data in near real-time to detect events such as canopy leaf-out. The second group of workflows are quality assurance filters, processing incoming sensor data through a series of criteria in order to produce "higher level" derived data products that are archived for use in post-hoc analyses. The third are for post-hoc analysis of data, representing a series of analyses and models that are used on archived sensor data and archived data from experimental treatments to assess the relative effects of fertilization and disease on competitive exclusion by the annual grasses described in section 2.1.

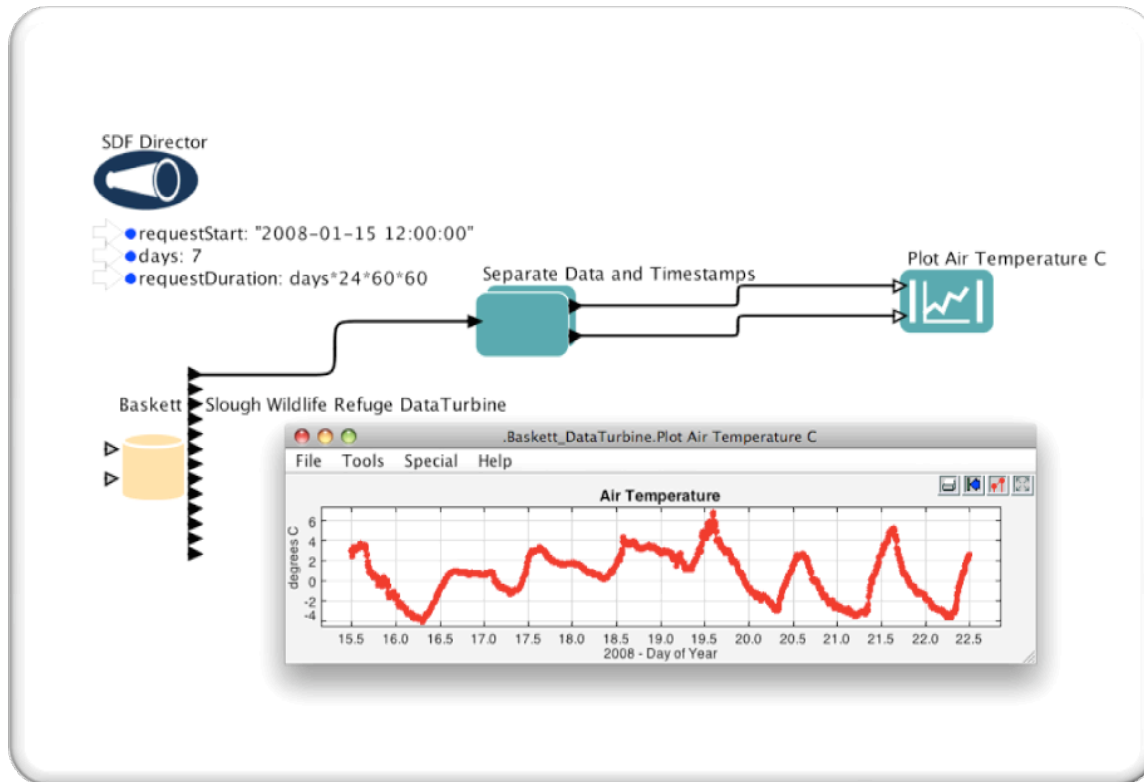


Figure2. Example Kepler workflow: Plotting sensor data from a DataTurbine server.

Many of these workflows require access to the sensor data that is being made available in DataTurbine. We developed a prototype DataTurbine sink actor within Kepler that exposes data from DataTurbine to downstream workflow components. This actor is configured by the user to connect to a DataTurbine server. The actor then automatically generates its output ports, each corresponding to a data channel from the DataTurbine server. Kepler workflow authors may then specify a time range of interest, connect the output ports of the DataTurbine actor to other actors, and operate on the received data within their workflows in whatever manner they see fit. A specific time range of data may be requested, or numerous such time ranges may be requested via iteration.

The simple workflow and its resulting plots in Figure 2 illustrate a workflow author requesting data for a specific time range (seven days starting at noon, Jan 15, 2008) using the DataTurbine actor, which has been configured by the user to point at our DataTurbine server. The channel requested, Air Temperature, is split into its data and timestamps components and then plotted.

DataTurbine also provides data request modes for streaming data in real-time, and support for these modes within Kepler is being developed. Streaming modes will be useful for "headless" (without a graphical user interface), continuously running ("batch mode") workflows that provide immediate notification when events of interest or problems occur.

## 5. Summary and Future Work

By developing Kepler actors that can consume data from sources such as DataTurbine, we provide tools useful to a broad set of data providers and consumers, and address needs beyond those specific to our use cases.

Based on our work with DataTurbine, we plan to develop more general Inspection, Monitoring, and Control APIs that work with other common sensor middleware software such as Boulder Real Time Technologies Antelope. With these interfaces, for example, a scientist will be able to browse data and be alerted when events of interest occur, and network engineers will be able to monitor deployed systems health and adjust data-stream rates.

As we develop new data-source actors for Kepler, we will unify and generalize those that already exist, leaving workflow authors with simpler sets of options from which to choose.

We plan workflows that format streaming sensor data into EML and then push these data as they progress through quality assurance and processing steps into a Metacat. For example, it will be possible for a network engineer to request from a Metacat Level 0 “raw” data, and a scientist a higher level “cleaned” data product.

We are closely following the Sensor Web Enablement initiative, and development efforts for being able to interact with implementations of the emerging OGC standards from within Kepler are also planned. For example, we will allow workflow authors a means of easily obtaining and using data from a Sensor Observations Service.

## References

- Altintas, I., C. Berkley, E. Jaeger, M. Jones, B. Ludäscher, S. Mock, 2004a. Kepler: An Extensible System for Design and Execution of Scientific Workflows. 16th International Conference on Scientific and Statistical Database Management. IEEE publication number P2146.
- Altintas, I., S. Bhagwanani, D. Buttler, S. Chandra, Z. Cheng, M. Coleman, T. Critchlow, A. Gupta, W. Han, L. Liu, B. Ludäscher, C. Pu, R. Moore, A. Shoshani, and M. Vouk, 2003. A Modeling and Execution Environment for Distributed Scientific Workflows, 15th Intl. Conference on Scientific and Statistical Database Management (SSDBM), Boston, Massachusetts.
- Altintas, I., A. Birnbaum, K. Baldrige, W. Sudholt, M. Miller, C. Amoreira, Y. Potier, and B. Ludäscher, 2005. A Framework for the Design and Reuse of Grid Workflows, Intl. Workshop on Scientific Applications on Grid Computing (SAG'04), LNCS 3458, Springer.
- Altintas, I., E. Jaeger, K. Lin, B. Ludäscher, and A. Memon, 2004b. A Web Service Composition and Deployment Framework for Scientific Workflows [abstract]. In: 2nd Intl. Conference on Web Services (ICWS), San Diego, California, July 2004.
- Baker, H. G. 1978. Invasion and replacement in Californian and neotropical grasslands, pp. 368-384 in J. R. Wilson, editor. Plant Relations in Pastures. CSIRO, East Melbourne.
- Borer, E. T., P. R. Hosseini, E. W. Seabloom, and A. P. Dobson, 2007. Pathogen-induced reversal of native dominance in a grassland community. *Proceedings of the National Academy of Sciences of the United States of America* 104:5473-5478.
- Botts, M., G. Percivall, C. Reed, and J. Davidson, 2007. OGC Sensor Web Enablement: Overview and High Level Architecture. OGC White Paper (OGC Document 07-165).
- Fegraus E.H., S. Andelman, M. B. Jones, M. Schildhauer, 2005. Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation. *Bulletin of the Ecological Society of America*: Vol. 86, No. 3 pp. 158–168.



- Halbert, S., and D. Voegtlin, 1995. Biology and taxonomy of vectors of barley yellow dwarf viruses. Pages 217-258 in C. J. D'Arcy and P. A. Burnett, editors. Barley Yellow Dwarf: 40 Years of Progress. The American Phytopathological Society, St. Paul, Minnesota.
- Jackson, L. E., 1985. Ecological Origins of Californias Mediterranean Grasses. *Journal of Biogeography* 12:349-361.
- Jones, M. B., C. Berkley, J. Bojilova, M. Schildhauer, 2001. Managing Scientific Metadata, *IEEE Internet Computing* 5(5) pp. 59-68.
- Ludäscher B., I. Altintas, C. Berkley, D. Higgins, E. Jaeger-Frank, M. Jones, E. Lee, J. Tao, Y. Zhao, 2006. Scientific Workflow Management and the Kepler System. Special Issue: Workflow in Grid Systems. *Concurrency and Computation: Practice & Experience* 18(10): 1039-1065.
- Tilak, S., P. Hubbard, M. Miller, and T. Fountain, 2007. The Ring Buffer Network Bus (RBNB) DataTurbine Streaming Data Middleware for Environmental Observing Systems. In *Proceedings of the Third IEEE international Conference on E-Science and Grid Computing (December 10 - 13, 2007)*. E-SCIENCE. IEEE Computer Society, Washington, DC, 125-133.

**MEETING THE CHALLENGES OF AN INTERNATIONAL, GRASSROOTS  
ORGANIZATION OF SITES DEPLOYING SENSOR NETWORKS: THE GLOBAL  
LAKE ECOLOGICAL OBSERVATORY NETWORK (GLEON)**

**B. J. Benson<sup>1</sup>, L. Winslow<sup>1</sup>, P. Arzberger<sup>2</sup>, C. C. Carey<sup>3</sup>, T. Fountain<sup>4</sup>, P. C. Hanson<sup>1</sup>, T. K. Kratz<sup>5</sup>, S. Tilak<sup>4</sup>**

<sup>1</sup>Center for Limnology, University of Wisconsin-Madison, 680 N. Park Street, Madison, WI 53706 USA; <sup>2</sup>California Institute for Telecommunications and Information Technology, and the Center for Research on Biological Systems, University of California-San Diego, La Jolla, CA 92093-0043 USA; <sup>3</sup> Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY 14853 USA; <sup>4</sup>San Diego Supercomputer Center, University of California-San Diego, La Jolla, CA 92093-0505 USA; <sup>5</sup>Trout Lake Station, Center for Limnology, University of Wisconsin-Madison, 10810 County Highway N, Boulder Junction, WI 54568 USA

**Abstract**

Realizing the full potential of embedded sensor networks to generate new scientific knowledge requires the sharing of data and expertise and the interdisciplinary collaboration of environmental scientists and information technologists and managers. The Global Lake Ecological Observatory Network (GLEON) is a grassroots network of limnologists, ecologists, information technology experts, and engineers with a common goal of building a scalable, persistent network of lake ecology observatories. GLEON's technological and organizational innovations provide models for how a grassroots organization can function to catalyze science based on environmental observing networks. Evolving solutions within GLEON to technological and organizational challenges include ways for sharing expertise, the development and deployment of software to enable effective management and sharing of sensor network data, the generation and documentation of GLEON operating principles and procedures, and the training of students in the new technology and large-scale scientific collaboration.

**Keywords:** sensor networks, cyberinfrastructure, ecoinformatics, environmental observatories, GLEON, lakes

**1. Introduction**

Many of the environmental challenges being addressed by scientists in today's world are regional or global in scope such as the effects of climate change, land use change, invasive species or human population growth and distribution. Often the study of these complex issues involves controls and interactions at multiple temporal and spatial scales. Embedded sensor networks are making new contributions to environmental sciences by extending the scales of spatial and temporal measurement (Estrin et al. 2003, Porter et al. 2005). Fully realizing the potential of this technology to generate new scientific knowledge will require the sharing of data and expertise and the interdisciplinary collaboration of environmental scientists and information technologists and managers.

A significant challenge then is finding effective ways to support these collaborations and scientific investigations that span regions and the globe. The Global Lake Ecological Observatory Network (GLEON) is a grassroots network of limnologists, ecologists, information technology (IT) experts, and engineers with a common goal of building a scalable, persistent

network of lake ecology observatories ([gleon.org](http://gleon.org); Kratz et al. 2006). Lakes, in particular, are a key ecosystem under stress in this changing world. The stated mission of GLEON is to facilitate interaction and collaboration among an international, multidisciplinary community of researchers focused on understanding, predicting, and communicating the impact of natural and anthropogenic influences on lake ecosystems by developing, deploying, and using networks of emerging observational system technologies and associated cyberinfrastructure. The inaugural GLEON meeting occurred in March 2005. As of April 2008, there were 298 individuals affiliated with GLEON, representing 31 countries. There have been six GLEON meetings attended by an international group of lake scientists and professionals from technical fields involved in sensor technology and information systems. Resources for building the GLEON community and its capacity for scientific collaborations have recently been significantly augmented by an NSF Research Coordination Network (RCN) grant. Technology development that is benefiting both GLEON and the Coral Reef Observatory Network (CREON) has been fostered through grants from the Gordon and Betty Moore Foundation.

We address here some evolving solutions to technological and organizational challenges faced by an international, grassroots organization of sites deploying sensor networks such as GLEON. These solutions include ways for sharing expertise, the development and deployment of software to enable effective management and sharing of sensor network data, the generation and documentation of GLEON operating principles and procedures, and the training of students in the new technology and large-scale scientific collaboration.

## **2. Methods and Techniques**

### **2.1 Sharing expertise.**

The sharing of expertise is an important benefit of a research network such as GLEON. The technology associated with sensor networks is a relatively new and rapidly evolving field. Multiple solutions and approaches exist within the diverse collection of sites. Communication at meetings, ongoing working groups, and the GLEON website all represent channels for sharing expertise. In addition, GLEON has developed the Lake Information Database ([gleon.org/lakes](http://gleon.org/lakes)), a web-accessible database of information about GLEON lakes and the sensors deployed on them. The displayed information includes an overall lake description, values for lake characteristics such as lake area and nutrient concentrations, a list of measurements being taken and the sensors that are used for these measurements. GLEON members enter information into the database through a web-enabled application that has both an administrative and user interface. The user interface includes the opportunity to add new vocabulary for measurement types and sensors as well as guidance text on entering information. User additions to the controlled vocabulary are then vetted by a subgroup of the GLEON Steering Committee.

### **2.2 IT development and deployment.**

Many GLEON sites have acquired or will be acquiring sensor technology and know how to deploy sensors and download data to a repository on a local computer. However, it is often the case that this repository, often a text file archive, is not easily shared, queried or made accessible via the Internet. To eliminate these gaps, we undertook the creation of information management system software to allow scientists to access the data via the Internet. This development has been supported by a number of synergistic grants and collaborations. To date, the software has been installed at four GLEON sites: Lake Erken in Sweden, Lake Sunapee in New Hampshire, Lake Annie in Florida, and the North Temperate Lakes LTER in Wisconsin.

These sites vary in the number of sensors deployed, the extent of legacy data, and the extent to which an information management system was already in place for non-sensor data. A team of people involved in deployment package development traveled to each site to install the system, and, in some cases, assisted with instrumentation deployment.

There were multiple components of the installation process (i.e., install, document, train, test, and evaluate) beyond the actual installation of the software that automated the data flow from downloaded text files to an Internet accessible database. For documentation, an installation report was prepared ([gleon.org](http://gleon.org)) for each site installation, and a repository of required technologies was maintained on the GLEON website. Local staff were given an overview of the technology and trained to change the system configuration and troubleshoot problems. System operation was tested under continuous operation conditions, individual component shutdown, and system reboot. The unique installation process at each site was evaluated.

An important part of the installation process is a site preparation component that is ideally generated by site personnel prior to installation. Documentation is required of the physical instrumented buoy system, the sampling regime, data download frequency and storage location, the vocabulary used for measurement, the logical hierarchy that allows the physical system to be represented in the data structure, a description of any legacy sensor data, and security constraints. The local site situation can generate additional requirements. For example, Lake Sunapee, which has a very active lake association, needs solutions for making data available in near-real time to the public; these solutions will be co-designed and implemented through a recently awarded NSF CI-Team grant.

Early on, it was recognized that to facilitate data sharing within GLEON, member sites without information management infrastructure would need to be brought up to speed. The requirements included managing high-resolution sensor network data and making those data web accessible through inexpensive or free tools and software. In addition, an emphasis was placed on ease of use and robustness as many GLEON members lack IT support. The envisioned deployment software package (Figure 1) consists of both off-the-shelf and custom software. The database is MySQL v.5 Community Server, and the database administrative tool is MySQL Administrator. Logger debriefing uses the Campbell LoggerNet (or PC208w). The data model (Vega; Winslow et al. 2008) provides storage flexibility, accommodating reconfigurations of the sensor network without changes to the database schema. During deployments Open Source DataTurbine ([www.dataturbine.org](http://www.dataturbine.org); Tilak et al. 2007) was tested. DataTurbine is open-source streaming data middleware that provides reliable data transport, a framework for integrating heterogeneous instruments, and a suite of services for data management, routing, synchronization, monitoring, and visualization. Inca ([inca.sdsc.edu](http://inca.sdsc.edu)) was tested to monitor the software and data management infrastructure and allow remote monitoring and troubleshooting. More technical details can be found in the installation reports ([gleon.org](http://gleon.org)).

### **2.3 Data sharing.**

The GLEON deployment package allows multiple destinations for the data stream, and in practice, the data from each of the deployment sites have been streamed to a central repository, in addition to the local repository. These data are then accessible through the GLEON web-site via custom query tools. The controlled vocabulary that is being developed for the Lake Information Database and the GLEON deployment package sets the stage for expanding data discovery and access beyond sites employing the deployment package. The controlled vocabulary includes measurements, sensors, and units.

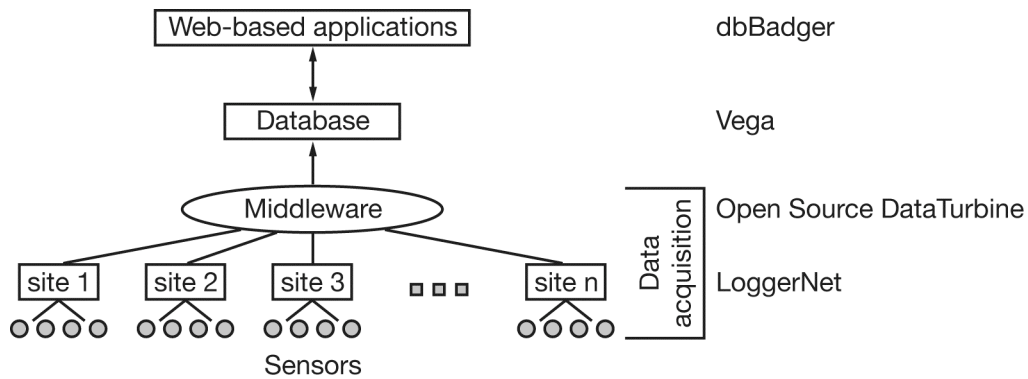


Figure 1. This diagram represents the data plane aspect of the system architecture for the GLEON deployment package. Data from sensors at distributed GLEON sites are routed to a centralized database as well as local databases through middleware. Web-based applications such as dbBadger allow users to query the database. The shared database is built on the Vega data model. Open Source DataTurbine has been tested at deployment sites for use as the middleware. System monitoring via Inca (not shown) pervades all the layers of the architecture and connects with components such as hardware (sensors and compute nodes), middleware, and databases.

### 3. Organizational issues

GLEON created operating principles and procedures to guide the growth and evolution of the network ([http://www.gleon.org/media/GLEON\\_OpPrincProc.pdf](http://www.gleon.org/media/GLEON_OpPrincProc.pdf)) as well as to have clear and transparent operating procedures. This document describes the organizational goals, values and principles, structure (including member and steering committee roles), and the policy on sharing of data. This document drew from the Pacific Rim Application and Grid Middleware Assembly (PRAGMA). This explicit statement of how to address these organizational issues increases organizational effectiveness, and the process of generating the structure and agreements has contributed to community cohesiveness.

### 4. Education

An important focus of the GLEON Research Coordination Network (RCN) is to inform, train, and mentor students while simultaneously preparing the next generation of scientists for large, collaborative, international, interdisciplinary science. The education of today's graduate students needs to prepare them to lead and collaborate within these larger, more complex research environments that are increasingly becoming a more extensive mode of conducting environmental science. Participation in GLEON provides several benefits to students: networking with other students and researchers across disciplines, learning new skill sets, the opportunity to experience a leadership role in an emerging organization, and traveling to different GLEON sites. Student activities have included conducting an informational meeting about GLEON for students attending the International Society for Limnology (SIL) meeting in Montreal in August 2007 and organizing the application process for students wanting to receive support to attend the GLEON VI meeting in Florida in February 2008 and the GLEON VII meeting in Sweden in September 2008. Fifteen students representing multiple disciplines and countries attended the GLEON VI meeting after a competitive application process. At GLEON VI, the students created the GLEON Student Travel-Funding Program, a cross-site collaborative effort for students to visit other GLEON sites and gain knowledge and experience within the network. This project is funded by the RCN.

## 5. Discussion

Grassroots networks such as GLEON can provide significant assistance to participating sites for implementing sensor networks and extending the technological development that supports sensor network deployment and information systems. Since its first meeting in 2005, GLEON has developed a web-accessible database of lake characteristics, measurements and sensor information for participating sites, developed and deployed a software suite that has enabled several sites to manage and share sensor data, conducted five more GLEON meetings at which extending technological developments that support the sites and network was a goal of the meetings. The community has benefited from relationships developed during the meetings, and there have been many instances of people with expertise from one GLEON site traveling to another site to assist in sensor deployment and various system design issues.

The data that are being shared within GLEON are pivotal to new scientific understanding. The sensor network measurements have potential to provide new estimates of ecosystem rates, better calibration of models, and identification of key controls over ecosystem process across multiple scales. They have illuminated the impact of events in near real-time such as the role of typhoons in restructuring lake ecosystems in a remote lake in the mountains of Taiwan (Tsai et al. 2008). Work is underway that capitalizes on the interplay between high frequency data and the development and extension of models of lake ecosystem processes. Now is an exciting time in which to explore the ways in which the new scales of measurement provided by sensor networks can expand the questions and models that scientists address, and GLEON is actively engaged in this exploration.

The grassroots organizational paradigm has contributed to the successes of GLEON (Hanson 2007). The openness of the organization to innovation by individuals and the ability to capitalize on heterogeneity across the sites, technologies, and scientific approaches are strengths of the grassroots approach. Considerable flexibility to foster multiple solutions to problems and the awareness of the importance of building trust among participants have both promoted cohesiveness within the GLEON community. The Coral Reef Ecological Observatory Network and the National Phenology Network are other examples of broad scale networks using a grassroots organizational paradigm.

Leaders and technology developers within GLEON are aware of multiple new challenges to be addressed. Installation of the deployment package for sensor networks can be streamlined so that it will become possible for the installation team to do remote deployments with the assistance of local staff. Sharing of real-time data streams is now possible for those sites using the deployment package. A future goal is to extend the portal for sharing data in ways that allow participation by the heterogeneous collection of information management systems that exist across the GLEON sites. GLEON members have expressed an interest in sharing a wider scope of data beyond the sensor network data, such as spatial information layers. An ongoing challenge is to find the resources to support new cyberinfrastructure development within GLEON and to facilitate participation in GLEON meetings and activities. As GLEON continues to grow, we will undoubtedly face additional challenges related to organizational structure and perhaps need to address issues related to an optimal network size.

## 6. Conclusions

GLEON's technological and organizational innovations provide models for how a grassroots organization can function to catalyze science based on environmental observing

networks, provide assistance to participating sites in implementing sensor networks, extend the technological development that supports sensor network deployment and information systems, and develop tools to promote sharing of expertise and data across a research network. Interdisciplinary partnerships of lake scientists, engineers, computer scientists, educators, and information technology and management experts are required to make the vision for GLEON a reality.

### **Acknowledgments**

We gratefully acknowledge support from the U.S. National Science Foundation through grants DBI-0639229, NEON 0446802, NEON 0446017, DEB-0217533, OCI 0627026 and OCI 0722067 and from the Gordon and Betty Moore Foundation.

### **References**

- Estrin, D., W. Michener, G. Bonito, and workshop participants. 2003. Environmental cyberinfrastructure needs for distributed sensor networks: a report from a National Science Foundation sponsored workshop. Scripps Institution of Oceanography, La Jolla, CA. 12-14 August 2003.
- Hanson, P. C. 2007. A grassroots approach to sensor and science networks. *Frontiers in Ecology and the Environment* 5(7): 343–343.
- Kratz, T. K., P. Arzberger, B. J. Benson, C. Chiu, K. Chiu, L. Ding, T. Fountain, D. Hamilton, P. C. Hanson, Y. H. Hu, F. Lin, D. F. McMullen, S. Tilak, and C. Wu. 2006. Towards a Global Lake Ecological Observatory Network. *Publications of the Karelian Institute* 145:51-63.
- Porter, J., P. Arzberger, H. Braun, P. Bryant, S. Gage, T. Hansen, P. Hanson, F. Lin, C. Lin, T. K. Kratz, W. Michener, S. Shapiro, and T. Williams. 2005. Wireless sensor networks for ecology. *Bioscience* 55:561-572.
- Tilak S., P. Hubbard, M. Miller, and T. Fountain. 2007. The Ring Buffer Network Bus (RBNB) DataTurbine streaming data middleware for environmental observing systems. e-Science 10/12/2007, Bangalore, India.
- Tsai, J.W., T.K. Kratz, et al. 2008. Seasonal dynamics and regulation of lake metabolism in a subtropical humic lake. *Freshwater Biology*. *In press*
- Winslow, L.A. et al. 2008. Vega: A flexible data model for environmental time series data. Environmental Information Management Conference 2008. Albuquerque, NM.

## CHALLENGES OF AVHRR VEGETATION DATA FOR REAL TIME APPLICATIONS

Molly E Brown

SSAI/NASA Goddard Space Flight Center

### Abstract

Remote sensing data has long been used to monitor global ecosystems for floods and droughts and AVHRR data, as one of the first product, has many users interested in receiving the data within hours of acquisition. With the introduction of a new series of sensors in 2000 (the AVHRR/3 series), the quality of the NDVI datasets available for real time environmental monitoring has declined. This paper provides evidence of problems of cloud contamination, calibration and noise in the real time data which are not present in the historical AVHRR NDVIg dataset. These differences introduce significant uncertainty in the use of the real time data, degrading their utility for detecting climate variations in near real time.

**Keywords:** AVHRR, NDVI, real time data, long term data record, data quality

### 1. Introduction

Remote sensing data used to monitor global ecosystems for floods and droughts have increased in importance in recent years. Increasing population density, industrialization and vulnerability to climate extremes has motivated the development of web-based geospatial decision support tools. These tools need accurate, reliable, synoptic information on environmental extremes. This information is often derived from remote sensing data. This paper focuses on the strengths, weaknesses, and opportunities posed by vegetation datasets developed for real time anomaly identification (Brown et al. 2006; van Leeuwen et al. 2006).

One of the first decision support tools to be developed using vegetation data was drought and flood monitoring in the context of famine early warning (Brown 2008). In the 1980s, vegetation data records were developed that provided information about the health of the plants over thousands of square kilometers simultaneously (Tucker 1979). Termed 'near real time datasets', these earth observation datasets are produced hours after data acquisition and are processed into image products and posted on the web for viewing by analysts from a variety of disciplines (Brown et al. 2007; van Leeuwen et al. 2006). Data from remote sensing is particularly useful in Africa, where other sources of data are less robust (Fensholt et al. 2006). Figure 1 shows the anomaly data from March, 2008 from the AVHRR sensor.

Like long term data records, these real time datasets must be self-consistent, calibrated and issues related to the remote sensing system need to be addressed. Just like all vegetation data products, most serious problem afflicting these datasets are clouds, which render any observation useless by obstructing the target, and to a lesser degree, effects of the bidirectional reflectance distribution function or BRDF (Los et al. 2000). This paper examines the success of these real time data products to balance the need for rapid delivery for applications that are time sensitive with processing that removes the effects of clouds and BRDF and other artifacts in the data. Most scientific investigations that use satellite-derived vegetation data (such as those for carbon modeling or climate change studies (Neigh et al. 2007; Slayback et al. 2003)), use the long term data record of normalized difference vegetation index (NDVI) produced six months after acquisition, which has dealt in a consistent way with both of these issues (Tucker et al. 2005).



This paper evaluates two real time vegetation datasets derived from visible and near infrared data from the Advanced Very High Resolution Radiometer (AVHRR) sensor. It compares the real time data to the much used long term data record from NASA's Global Inventory Monitoring and Mapping Systems (GIMMS) group NDVIg vegetation dataset from the same period. For comparison, real time data from the French sensor SPOT-Vegetation (SPOT-Vegetation 2004) will also be examined, although the products from this sensor are only available several days after acquisition.

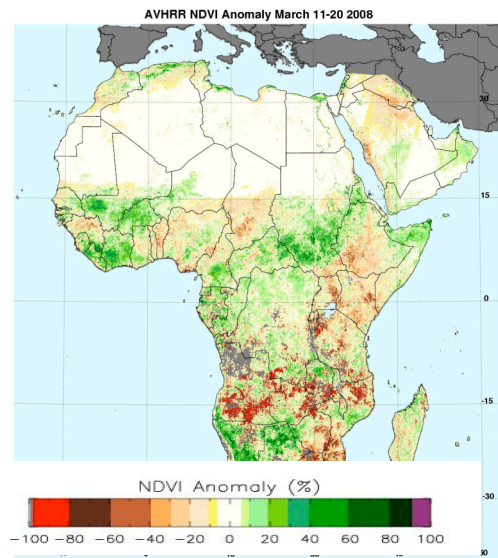


Figure 1. AVHRR percent anomaly  
March 11-20, 2008

## 2. Data and Methods

Data used in this study include the AVHRR NDVIg product (G), the real-time G product (RG), the NOAA-17 real time product (N17), and the S10 SPOT Vegetation data (SP) reprojected and regridded to match the Albers 8km resolution of the GIMMS data. The following section describes the datasets and then the methods will be described.

### 2.1 Data products

The GIMMS NDVIg Historical dataset is a 15 day, maximum value AVHRR normalized difference vegetation data composites (Holben 1986) from the NASA Global Inventory Monitoring and Modeling Systems (GIMMS) group at the Laboratory for Terrestrial Physics (Tucker et al. 2005) from July 1981 to December 2006. The GIMMS operational dataset incorporates data from sensors aboard NOAA-7 through 14 with the data from the AVHRR on NOAA-16 and 17 using SPOT data as a bridge for a by-pixel inter-calibration. After calibration, the AVHRR NDVI data dynamic range was adjusted to values of  $-0.05$  to  $0.95$  to match more closely that of the SPOT- and MODIS-based NDVI. In addition, the NDVIg data has had an algorithm applied that replaces pixels with missing data, data obscured by clouds and data with depressed values due to significant sub-pixel cloud contamination with interpolated data.

The AVHRR real time datasets include the RG and the N17 datasets. The RG real time NDVI product uses the same code as was used to produce the NDVIg, but with a calibration computed approximately once every six months. The RG product has had the inter-calibration of the datasets applied and the adjusted dynamic range of the NDVIg dataset, but has not had the interpolation routine removing clouds. Cloud detection is a channel 5 (T5) temperature threshold technique, using 285 degrees Kelvin for Africa to detect pixels with low temperatures indicating cold cloud tops.

The N17 data is a NOAA-17-only dataset processed using the imbedded information present in NOAA-17 level 1b AVHRR data. This release uses three methods to detect daytime clouds over land, including a T5 temperature threshold of 285 degrees Kelvin over Africa, the cloud, snow and cloud shadow method (Saunders and Kriebel 1988) as implemented in the SPOT Vegetation dataset and the NOAA CIMSS Level-1b (v4) CLAVR-x method (Heidinger et

al. 2006). The navigation has been improved by correcting the registration offset between ascending/descending nodes and GAC Earth location interpolation.

The SPOT Vegetation data used in this study are VGT-S10 (ten day synthesis) products. The ‘S10-composited’ data (spectral band data, data quality and NDVI) covering the period May-1998 to June-2004 were acquired for analysis. Post-processing includes reprojection from the native global Mercator to a continental Albers projection, regriding to 8km resolution, regional sub-setting, cloud screening, and land masking (Brown et al. 2006). The SPOT data was processed in collaboration between the USDA's Foreign Agricultural Service (FAS) Production and Crop Assessments Division (PECAD) and NASA/GSFC's GIMMS group.

## 2.2 Methods

To compare the information from multiple datasets, I examined both time series extracted from the data and data from the entire continent of Africa from September 2002 to December 2007. Table 1 lists the locations where time series data were extracted and examined. I subtracted time series from each other to determine the differences between the information in the real time datasets and the NDVIg datasets, and I compared the ability of these real time datasets from AVHRR to identify periods of flood and drought to that of the SPOT data. I used also the mean, standard deviation and variance of the continental data as a measure of its stability.

City	Country	Region	Latitude	Longitude
Bonkougou	Niger	West	14.04	3.22
Louga	Senegal	West	15.63	-16.17
Kano	Nigeria	West	11.89	8.53
Mongo	Chad	West	12.12	18.64
Malakal	Sudan	East	9.53	31.65
Goba	Ethiopia	East	4.74	39.30
Baydhabo	Somalia	East	3.12	43.64
Dodoma	Tanzania	South	-6.09	35.71
Mbandaka	DRC	Central	0.05	18.26
Messina	South Africa	South	-22.25	30.09
Dutlwe	Botswana	South	-23.98	23.90
Tsumeb	Namibia	South	-19.25	17.71

Table 1. Locations where time series were extracted in Africa

## 3. Results

Figure 2 shows the difference between the GIMMS NDVIg and the two AVHRR real time products, the N17 and the RG, period by period. Several issues of the real time datasets can be noted from the figure: systematic cloud contamination, calibration issues across sensors and noise. The difference between the real time and historical NDVIg dataset in some regions varies

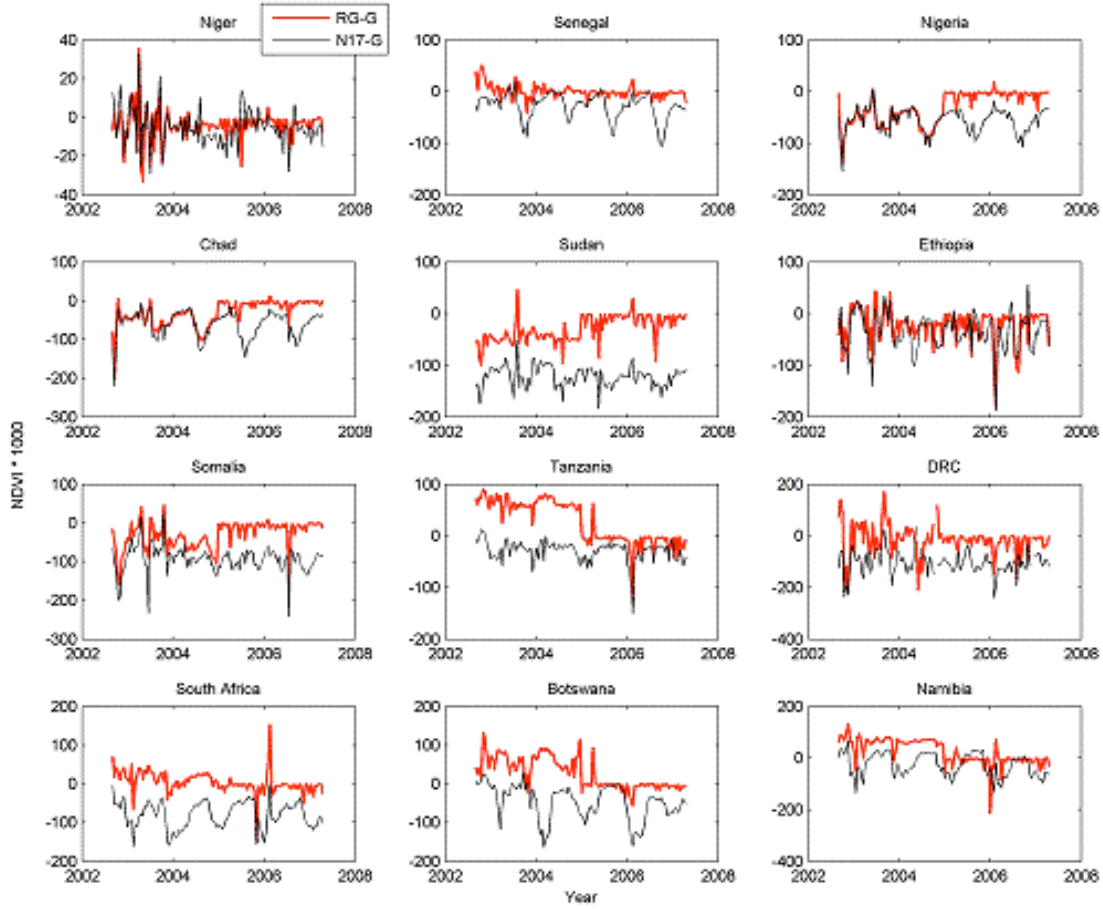


Figure 2. Time series of AVHRR real time datasets subtracted from the AVHRR GIMMS NDVIg dataset

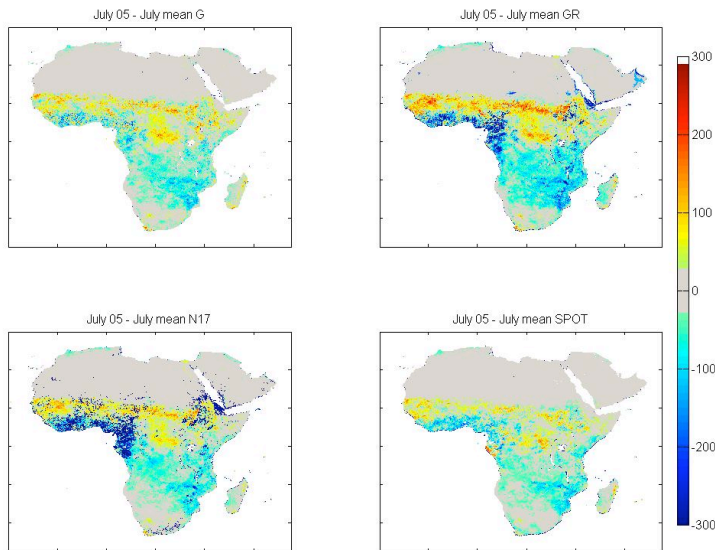


Figure 3. July 2005 anomaly from mean data from 2002-2007, showing difference between products for the same period.

seasonally due to cloud contamination which affects the real time data more than the post-processed, corrected NDVIg data. Large differences in calibration can also be seen which change through time, with significant shifts in 2004 due to the change from NOAA-16 to NOAA-17 data in the RG product. The N17 product only uses data from NOAA-17. The RG product also has significantly more high frequency noise than the N17 product.

Figure 3 shows the varying continental-scale anomalies resulting in the different processing schemes of the real time datasets, even for the same period. Although the overall pattern of the anomaly is similar, with the exception of the cloud-induced negative anomaly focused on the Guinea coast which varies in size depending on the product, the size of the anomaly varies by product. The anomalies are nearly twice in the GR product than the G anomaly and the SPOT anomaly. The N17 anomaly is slightly larger than the G anomaly, but still significantly cloud contaminated.

#### **4. Discussion**

AVHRR's wide spectral bands cause the data to be quite sensitive to water vapor in the atmosphere. Increases in water vapor results in lower NDVI signal, which can be interpreted as an actual change if no correction is applied (Pinheiro et al. 2004). The maximum value composite lessens these artifacts (Brown et al. 2006), but issues obviously still remain. This study shows that when comparing cloud-corrected NDVIg data to real-time data without significant post-processing cloud removal causes significant false negative anomalies. These false negative anomalies cause difficulty for operational data users seeking to use AVHRR data to monitor widespread changes in food availability or pasture for animals. As the historical NDVIg data has improved in stability, completeness, and utility for scientific trend analysis, it has become much more difficult to match in real time processing as the techniques used to correct the data do not do well at the ends of time series (such as decomposition techniques).

Efforts by the GIMMS group to produce a new real time data product, such as the N17 product, have resulted in a much better calibrated product, but one with significant differences from the historical dataset in cloud detection. Although the N17 product detects more clouds and flags them than the historical data, the result is a significant decline in the amount of data available for trend detection. The quality of the product then suffers, as much of the area of interest is obscured behind cloud masks as Figure 3 shows.

The SPOT Vegetation data product is far superior to either of the AVHRR data products due to a much higher data density due to its 1km resolution and to its superior cloud detection. SPOT data has had significant problems with calibration, however, which have reduced the mean productivity in Africa making for anomaly products that are higher than they should be.

#### **5. Conclusions**

The reality of user perceptions of declining quality of real-time AVHRR data since the introduction of the AVHRR/3 sensor has been demonstrated by this analysis. The GIMMS group had to greatly increase the level of complexity of the calibration, navigation and cloud removal techniques in order to incorporate data from both the AVHRR/2 NOAA-7 through 14 with the AVHRR/3 series of NOAA-16 and 17. This complexity has severely impacted the ability to implement a reasonable process on data that must be delivered within hours of acquisition.

Despite over two and a half decades of research and analysis and a high degree of similarity between the AVHRR/3 sensors, calibration and cloud detection issues in real time data

still remain. These problems are land-cover dependent, with some regions showing much better similarity between the real time and historical NDVIg dataset than others. As we move into the era where the replacement for MODIS will be a sensor with yet another and far more complicated sensor design (VIIRS), and where AVHRR data will not be available, we can anticipate serious problems bringing multiple datasets together for critical climate change studies when the sensors from which the data were derived are completely different. Thus this research shows that continued investment in research on AVHRR data integration with MODIS and VIIRS datasets is essential, as well as the continuation of the AVHRR sensor to ensure data continuity if the research does not come to fruition.

## References

- Brown ME (2008) *Famine Early Warning Systems and Remote Sensing Data*. Springer Verlag, Heidelberg, 357 pp.
- Brown ME, Funk C, Galu G and Choularton R (2007) Earlier Famine Warning Possible Using Remote Sensing and Models. *EOS Transactions of the American Geophysical Union*, 88: 381-382.
- Brown ME, Pinzon JE, Didan K, Morisette JT and Tucker CJ (2006) Evaluation of the consistency of long-term NDVI time series derived from AVHRR, SPOT-Vegetation, SeaWiFS, MODIS and LandsAT ETM+. *IEEE Transactions Geoscience and Remote Sensing*, 44: 1787-1793.
- Fensholt R, Nielsen TT and Stisen S (2006) Evaluation of AVHRR PAL and GIMMS 10-day composite NDVI time series products using SPOT-4 vegetation data for the African continent. *International Journal of Remote Sensing*, 27: 2719-1733.
- Heidinger A, Baum BA and Yang P (2006) Consistency of cloud ice properties estimated from MODIS, AVHRR and SEVIRI, Radiative properties of clouds (Joint Session with 12th Conference on Atmospheric Radiation & 12th Conference on Cloud Physics), Madison, WI.
- Holben B (1986) Characteristics of Maximum-Value Composite Images from Temporal AVHRR Data. *International Journal of Remote Sensing*, 7: 1417-1434.
- Los SO, Collatz GJ, Sellers PJ, Malmstrom CM, Pollack NH, DeFries RS, Bounoua L, Parris MT, Tucker CJ and Dazlich DA (2000) A Global 9-year Biophysical Land Surface Dataset from NOAA AVHRR data. *Journal of Hydrometeorology*, 1: 183-199.
- Neigh CSR, Tucker CJ and Townshend JRG (2007) Synchronous NDVI and Surface Air Temperature Trends in Newfoundland: 1982-2003. *International Journal of Remote Sensing*, 28: 2581-2598.
- Pinheiro AC, Privette JL, Mahoney R and Tucker CJ (2004) Directional Effects in a Daily AVHRR Land Surface Temperature Data Set over Africa. *IEEE Transactions in Geoscience and Remote Sensing*, 42: 1941-1954.
- Saunders RW and Kriebel KT (1988) An improved method for detecting clear sky and cloudy radiances from AVHRR data. *International Journal of Remote Sensing*, 9: 123-150.
- Slayback DA, Pinzon JE, Los SO and Tucker CJ (2003) Northern hemisphere photosynthetic trends 1982-99. *Global Change Biology*, 9: 1-15.
- SPOT-Vegetation (2004) *Vegetation Program Web Site*. VITO Belgium.
- Tucker CJ (1979) Red and Photographic Infrared Linear Combinations for Monitoring Vegetation. *Remote Sensing of Environment*, 8: 127-150.

- Tucker CJ, Pinzon JE, Brown ME, Slayback D, Pak EW, Mahoney R, Vermote E and Saleous N (2005) An Extended AVHRR 8-km NDVI Data Set Compatible with MODIS and SPOT Vegetation NDVI Data. *International Journal of Remote Sensing*, 26: 4485-4498.
- van Leeuwen W, Orr BJ, Marsh SE and Herrmann SM (2006) Multi-sensor NDVI data continuity: Uncertainties and implications for vegetation monitoring applications. *Remote Sensing of Environment*, 100: 67-81.

## **EVOLVING SENSOR WEB PROTOCOLS FOR SENSOR DATA MANAGEMENT**

**Helen Conover<sup>1</sup>, Kathryn Regner<sup>1</sup>, Manil Maskey<sup>1</sup>, Jessica Lu<sup>1</sup>, Xiang Li<sup>1</sup>, H. Michael Goodman<sup>2</sup>**

<sup>1</sup>University of Alabama in Huntsville

<sup>2</sup>National Aeronautics and Space Administration, Marshall Space Flight Center

### **Abstract**

Standard interfaces for data and information access facilitate data management and usability by minimizing the effort required to acquire, catalog and integrate data from a variety of sources. The authors have prototyped several data management and analysis applications using Sensor Web Enablement Services, a suite of service protocols being developed by the Open Geospatial Consortium specifically for handling sensor data in near real time. This paper provides a brief overview of some of the service protocols and describes how they are used in three different sensor web projects involving near real time management of sensor data.

**Keywords:** sensor web, standards, geospatial data, near real time

### **1. Introduction**

One of the challenges facing today's Earth system scientist is being able to navigate the ever increasing amount of sensor data available from a wide variety of in situ and dynamic environmental sensors. A key aspect of managing these sensor data in near real time is providing efficient discovery, access and processing of sensor observations. The Open Geospatial Consortium (OGC) suite of Sensor Web Enablement (SWE) specifications, some newly released and others under development, provides standards for data and information acquisition from sensor systems and data repositories. The OGC is an international consortium of industry, academic and government organizations using a voluntary consensus process to collaboratively develop open standards for geospatial data and information services. The OGC SWE standards framework provides specifications for interfaces, protocols and encodings that are designed to enable implementation of interoperable, service-oriented networks of sensors and applications (Botts et al. 2007). Providing such standard interfaces to sensor data can minimize the custom software required for management, visualization and analysis of different types of sensors and observations. The authors have implemented several types of SWE services for selected sensor data sources, then combined these services in different ways to prototype a variety of data processing and management applications, including weather forecasting and mission monitoring. This paper describes these efforts to explore the readiness of emerging SWE standards to integrate both Earth observations and forecast model output into new data acquisition, assimilation and management strategies.

### **2. Sensor Web Enablement Services**

Sensor web enablement services implemented for this study include Sensor Observation Services (SOS) and Sensor Alert Services (SAS). SOS provides a web service interface for requesting, filtering and retrieving sensor system information and observations, while SAS provides a web service interface for advertising, publishing and subscribing to alerts from sensors. Other SWE protocols used in these prototypes include the Observations and

Measurements Schema (O&M), an XML schema for encoding sensor data objects, and Sensor Model Language (SensorML), an XML schema for describing a functional model of a sensor system and related processes. Documentation for all approved OGC standards is freely available at <http://www.opengeospatial.org/standards/>.

## 2.1 Sensor Observation Services

The SOS protocol provides a standard interface for requesting, filtering and retrieving sensor system information and observations. Sensor Observation Services are typically Representational State Transfer (REST) style web services which allow the user or calling program to select any number of the observation variables available from the data source, and to subset the data by spatial and temporal range, thereby significantly reducing data volume and transfer time. Sensor data is converted from its native format to O&M and may be delivered as either ASCII or binary attachment, in order to foster data/application interoperability and interuse of multiple sensor data products. The same service can provide access to archived data, or to near real time data streams as they are acquired at our data repository.

The authors and their colleagues in several research projects have developed SOS data access services for observations from sensors on a variety of platforms, including ocean buoys, ground stations, aircraft and satellites. This activity has been complicated by the fact that the SOS specification was still evolving rapidly until its approval as an OGC standard in October 2007. At this writing, the SOS clients and servers developed for our projects by the authors and multiple partners are still being upgraded to SOS version 1.0, leading to version mismatches in some distributed applications.

## 2.2 Sensor Alert Service

The SAS protocol provides a web service interface for advertising, publishing and subscribing to alerts from sensors. A Sensor Alert Service is not an event notification system, rather, it is a registry that cross-references the different types of alerts available from a given sensor system and the consumers subscribing to these alerts. Users send subscription requests to the SAS, which returns a communication endpoint for the alert subscribed to. The user must then open a connection to the communication endpoint to receive alerts from the sensor system. For this study, the authors are leveraging the SAS package from 52°North (<http://52north.org/>), a company which promotes the development and application of free open source geo-software for research, education, training and practical use. This SAS implementation uses Extensible Messaging and Presence Protocol (XMPP), an open XML technology for presence and real time communication used in applications such as instant messaging (Saint-Andre 2005).

The SAS specification is not yet an approved standard, but still under OGC community review. However, the authors were able to obtain the latest “snapshot” release of the SAS package from 52°North, which follows the current OGC SAS specification. Several modifications were needed for use with our prototype sensor systems, including:

- Modifications to make it possible to send geographical information (i.e., bounding box) in an alert.
- Added a “DescribeAlert” operation to provide the message structure of an alert.
- Developed an *SAS client* which provides a “publishAlert” operation for sensors to call to publish alerts programmatically rather than manually.



- Developed an *SAS-SOS adapter* which will query an SOS for data on selected alert conditions.

### 3. Sensor Observation Service Registry

A service registry contains information including descriptions of sensors and their observations to help users locate data and services to meet their needs. The authors have implemented an SOS registry to provide for registration and easy discovery of an increasing number of SOSs being implemented across multiple projects. The SOS registry provides two service Application Programming Interfaces (APIs) which permit data providers to register their sites and observation offerings, and end-users or automated services to search the registry for observation offerings and/or sensors having characteristics that meet their criteria. This service registry is currently being used in two projects, SMART and OOSTethys, described in section 4. The web interfaces to these services are available at <http://smart.uah.edu/catalog/> and <http://score.itsc.uah.edu/MMI/>, respectively.

The “SetCapabilities” service records the observation offerings or updates the observation offerings of an SOS. The service accepts a valid “GetCapabilities” URL for the SOS and harvests observation offering and sensor metadata from the “GetCapabilities” response. Human approval is needed for registration verification to prevent any unsupported registration. The registry also updates information about the SOSs by frequently polling the registered SOSs. The “GetURLs” service provides a means to discover registered SOSs. This is a simple REST service method that accepts various parameters and responds with information about the SOS provider, observation offerings, and sensors.

Although our current implementation of the SOS registry does not follow any standard specification for a catalog search API, we are investigating use of Catalogue Services for the Web (CSW) as a standard interface for a more robust SOS registry (Maskey 2008).

### 4. Sensor Prototypes

The authors are using SWE services for three prototype projects requiring near real time access to sensor data: Sensor Management Applied Research Technologies (SMART) On Demand Modeling (ODM), Real Time Mission Monitor (RTMM), and OOSTethys.

#### 4.1 SMART-ODM:

SMART (<http://smart.uah.edu>) is working with NASA’s Short-term Prediction Research and Transition (SPoRT) Center to develop a sensor web-enabled processing workflow to intelligently assimilate Atmospheric Infrared Sounding (AIRS) satellite temperature and moisture retrievals into a regional Weather Research and Forecast (WRF) model over the southeastern United States (Goodman et al. 2007). At SPoRT, a North American Mesoscale (NAM) forecast is used as the initial conditions for a regional WRF model run. The addition of current weather observations (such as those from AIRS) to the initial conditions can improve the accuracy of a WRF forecast, but assimilating voluminous satellite data is computationally expensive. Modelers and IT experts on the SMART team have worked together closely to determine how sensor web-enabled data access and analysis tools can best facilitate near real time data assimilation decisions. The SMART workflow, shown in Figure 1, involves mining NAM forecasts for interesting weather phenomena, then determining whether AIRS observations are coincident with the detected weather events. The assumption is that assimilating AIRS observations of anomalous weather conditions will improve the forecast.

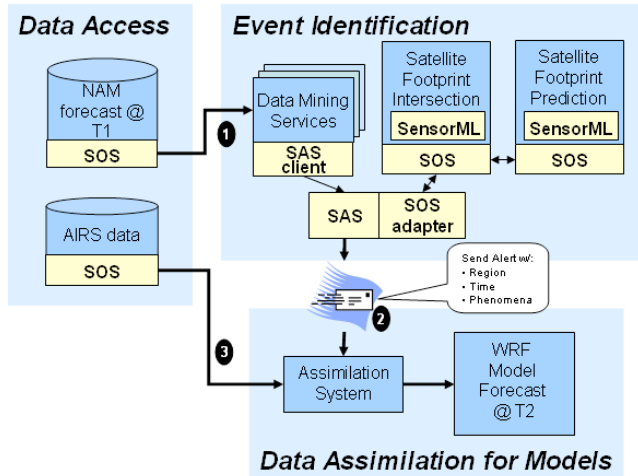


Figure 1: SMART satellite data assimilation for models using sensor web services

SMART uses data mining services (Graves et al. 2007) to provide the intelligence for automated decision making along with a variety of SWE protocols for data access and alert services, and for process chain definition. *Sensor Observation Services* provide web access to both NAM forecasts and AIRS observations. SOSs for Satellite Footprint Prediction and Intersection use *SensorML* models and process chains to determine satellite locations at any given time, and whether a given instrument footprint intersects a specified spatio-temporal region of interest. Finally, a *Sensor Alert Service* notifies subscribers when sensor

observations will be coincident with forecast weather events. Receipt of such an alert initiates satellite data assimilation for subsequent forecasts.

#### 4.2 RTMM:

The NASA Real Time Mission Monitor (RTMM) (<http://rtmm.nsstc.nasa.gov/>) is an interactive field experiment asset management and data visualization tool that incorporates SWE tools and protocols (Blakeslee et al. 2007). RTMM has been used in the NASA African Monsoon Multidisciplinary Analyses (NAMMA), Tropical Composition, Clouds and Climate Coupling (TC4) and Arctic Research of the Composition of the Troposphere from Aircraft and Satellites (ARCTAS) field experiments. This situational awareness tool, shown in Figure 2, integrates satellite, airborne and surface data sets; weather information; model and forecast outputs; and vehicle state data (e.g., aircraft navigation, satellite tracks and instrument field-of-views) for managing field experiments. The goal of the RTMM is to optimize science and logistic decision-making during field experiments by presenting timely data and graphics to the users to improve real time situational awareness of the experiment's assets. RTMM is evolving towards a more flexible and dynamic combination of sensor ingest, network computing, and decision-making activities through the use of a service oriented architecture based on community standards and protocols, such as SOS.

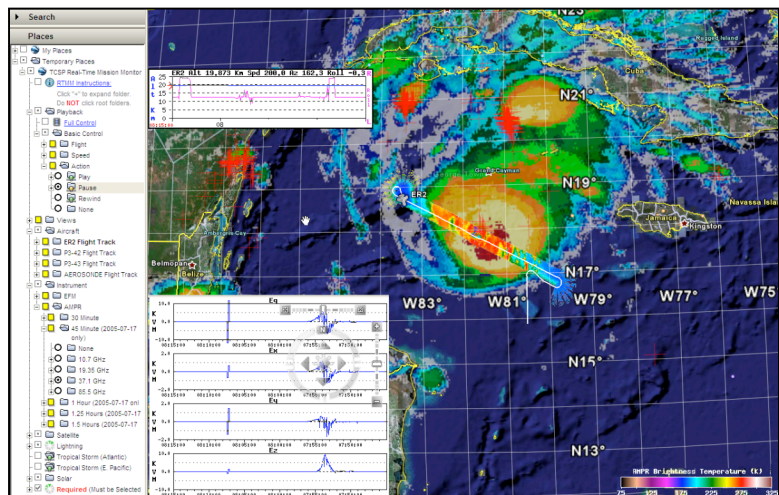


Figure 2: NASA aircraft flight track and sensor data over hurricane Emily, displayed on Google Earth with Real Time Mission Monitor

Currently, RTMM uses the Satellite Footprint Prediction SOS

to provide accurate information on the current location of Earth observing satellites, as well as to help mission planners prepare aircraft flight plans coincident with future satellite overpasses. In the future, SWE protocols can be used to communicate with mission sensors to access and display data in real time, and potentially to task sensors for specific observations.

### 4.3 OOSTethys and the Ocean Science Interoperability Experiment:

#### OOSTethys

(<http://www.oostethys.org/>) is a collaborative project in which members of the ocean science community are using OGC standards to prototype an ocean observing "system of systems," Figure 3. OOSTethys partners develop, test and document reference implementations of OGC-compliant software, and have created a working prototype of networked, interoperable, real time data systems. The related Ocean Science Interoperability Experiment (OCEANS IE) is an OGC initiative with a current goal of comparing the OGC Sensor Observations Service and Web Feature Service (WFS) protocols

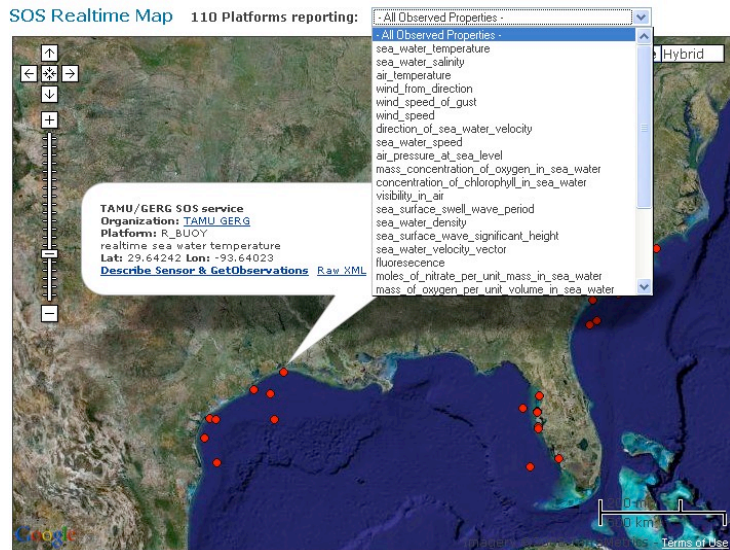


Figure 3: Data from OOSTethys ocean buoy SOSs displayed on Google Map

as applied to ocean data in a variety of data formats including text files, netCDF files, relational databases, and possibly native sensor output. One important outcome of this project has been the "cookbooks" and "how-to" documents that simplify the implementation and installation of SOS.

### 5. Conclusions

The technologies described here, which provide standard interfaces to sensor systems, can serve as the basis for a complete near real time data management system, including sensor systems, data repositories, and registries of both sensors and observations. Standard interfaces for data and information access will improve data usability by minimizing the effort required to discover and integrate new data sources into a scientific investigation or decision process. Each of the prototype systems described makes use of sensor observations acquired in near real time via Sensor Observation Services. Such systems could also task sensors to acquire additional observations of specific regions or phenomena, based on alerts received or issued by the system.

While OGC SWE technologies are new and evolving, reference implementations and "cookbooks" for many of the services are freely available. Science/IT collaboration is critical to the implementation of systems using these and other advanced technologies. That is, a team comprising both scientists and software engineers will result in a more scientifically viable, real world result than a team of only scientists or only software engineers. In the future, the results of examining and evolving these concepts will result in better exploitation of sensor observations for not only researchers and decision makers but for the benefit of casual end users as well.

## Acknowledgements

The SMART and RTMM teams consist of Rich Blakeslee, Gary Jedlovec, Michael Goodman and Robbie Hood of NASA/MSFC, and Gregoire Berthiau, Mike Botts, Helen Conover, Tony Cook, John Hall, Matt He, Xiang Li, Jessica Lu, Manil Maskey, Kathryn Regner, and Brad Zavodsky of UAH. The authors would also like to thank the NASA/MSFC SPoRT Center for assistance in designing the intelligent data assimilation prototype, and Johannes Echterhoff and Jan Torben Heuer of 52°North for insight into the 52°North approach to SWE services and access to the latest SAS software.

The primary support for this research is the Advanced Information Systems Technology program, sponsored by NASA's Earth Science Technology Office. This work builds on previous and concurrent research sponsored by other NOAA and NASA programs.

## References

- Blakeslee, R., Hall, J., Goodman, H. M., Parker, P., Freudinger, L., and He, M., 2007. The Real Time Mission Monitor - A Situational Awareness Tool for Managing Experiment Assets. NASA Science Technology Conference. College Park, MD.
- Botts, M., Percival, G., Reed, C., and Davidson, J., 2007. OGC Sensor Web Enablement: Overview and High Level Architecture. OpenGIS White Paper: OGC 07-165.
- Goodman, H. M., Jedlovec, G., Conover, H., Botts, M., Robin, A., Blakeslee, R., Hood, R., Ingenthron, S., Li, X., Maskey, M., and Stephens, K., 2007. The Sensor Management for Applied Research Technologies (SMART) Project. NASA Science Technology Conference. University of Maryland.
- Graves, S., Ramachandran, R., Keiser, K., Maskey, M., Lynnes, C., and Pham, L., 2007. Deployable Suite of Data Mining Web Services for Online Science Data Repositories. 87th AMS Annual Meeting. San Antonio, TX.
- Maskey, M., cited 2008. Issues related to using degree CSW for SOS registry. [online at <http://oostethys-trac.tamu.edu/oostethys/attachment/ticket/24/degree-csw-issues.doc>.]
- Saint-Andre, P., 2005. Streaming XML with Jabber/XMPP. IEEE Internet Computing 9(5): 82-89.

**INTEGRATING ECOLOGICAL DATA:  
NOTES FROM THE GRASSLANDS ANPP DATA INTEGRATION  
PROJECT**

**Judith B. Cushing<sup>1</sup>, Nicole E. Kaplan<sup>2</sup>, Christine Laney<sup>3</sup>, Juli Mallett<sup>1</sup>,  
Ken Ramsey<sup>4</sup>, Kristin Vanderbilt<sup>5</sup>, Lee Zeman<sup>1</sup>  
Jincheng Gao<sup>6</sup>, Judith Kruger<sup>7</sup>, Carri LeRoy<sup>1</sup>, Daniel Milchunas<sup>8</sup>, Esteban Muldavin<sup>5</sup>**

<sup>1</sup>The Evergreen State College, Olympia, WA 98505, <sup>2</sup>Department of Soil and Crop Sciences, Colorado State University, Fort Collins, CO 80523, <sup>3</sup>Jornada Experimental Range, New Mexico State University, Las Cruces, New Mexico 88003, <sup>4</sup>Department of Biology New Mexico State University, Las Cruces, New Mexico 88003, <sup>5</sup>Department of Biology, University of New Mexico, Albuquerque, New Mexico, 87131, <sup>6</sup>Department of Biology, Kansas State University, Manhattan, KS 66506, <sup>7</sup>South African National Parks, Scientific Services, Skukuza, South Africa, 1350, <sup>8</sup>Department of Forest, Rangeland, & Watershed Stewardship, Colorado State University, Fort Collins, CO 80523

**Abstract**

Trends in annual aboveground net primary productivity (ANPP) at regional and global scales are an important component of the structure and function of ecosystems across spatial and temporal gradients in a changing world. Ecologists are interested in conducting cross-site or large-scale integration and analysis of annual ANPP values, but are often hindered by the lack of standard methodologies for data collection, data management practices and detailed metadata documentation across sites. The Grasslands ANPP Data Integration (GDI) project has brought together experts in ecology, information management, and computer science to address the challenges of integrating ANPP data. Together, we have created a centralized database of annual ANPP data and metadata from five national and international Long Term Ecological Research (LTER) grassland sites. The database contains ANPP data at a level of granularity appropriate to each site, but standardizes vegetation species codes and sampling location metadata to facilitate cross-site comparison. This approach is important to local ecologists and information managers as no data are lost, and data can still be aggregated to the proper level of granularity for statistically valid cross-site analysis. The GDI database facilitates transformation, integration, and exploration of site-specific ANPP data, and preliminary cross-site statistical analyses and synthetic research. The GDI team has created processes and tools that will enable future warehousing of ANPP data by streamlining data insertion, update, integration, and standard metadata documentation and species information. This paper presents a description of the GDI data model, data transformation and integration techniques, and quality assurance standards. Lessons learned that might be applicable to other ecological and scientific data integration are also included.

**Keywords:** Ecological informatics, database integration, ecological synthesis, ANPP, biotic data semantics

**1. Introduction**

The GDI (Grasslands ANPP Data Integration) project is a joint effort among ecologists interested in annual aboveground net primary production (ANPP), Long Term Ecological Research (LTER) Information Managers, and computer scientists interested in data integration

and semantics. ANPP datasets represent core areas of research in many programs, including the Long Term Ecological Research Network,<sup>1</sup> Oak Ridge National Laboratory,<sup>2</sup> and the Global Terrestrial Observing System,<sup>3</sup> and are an important measurement for assessing ecosystem structure and function, biodiversity and ecosystem services, including carbon sequestration (Parton et al. 1995).

The influence of changing climate on ANPP is a question of great interest to ecologists. Knapp and Smith (2001) assessed the temporal dynamics of ANPP across eleven LTER sites in the United States, and suggested that grassland ANPP will be very responsive to future climatic changes. Their analyses were conducted with total annual ANPP values from each site. More refined analyses of ANPP values might also be of interest; for example, ANPP broken down by plant species and life forms would help assess community and population responses to variability in precipitation and help predict how grasslands might respond to global change phenomena. Synthesizing long-term datasets of species or life form level ANPP data from different regions and ecosystems is a critical first step towards conducting this research. However, ecologists and information managers across different LTER sites have in the past experienced challenges to integrating ANPP data from multiple sources.

Given the importance of ANPP to the ecological community, computer scientists, information managers, and ecologists both within and peripherally connected to the LTER and ILTER networks initiated a project to integrate ANPP data from several sites so that synthesis research across sites could be more easily conducted. Project objectives were to make the integration process more efficient, enable cross-site analysis, conserve fine levels of data granularity, and eventually accommodate ANPP data from sites outside the grassland biome, as well as other grassland sites. A reliable and useful integrated data product requires documenting the data as they are loaded, determining a statistically valid level of comparison, and transforming the data into a standardized format. A long term sustainable data warehouse, however, is not feasible without semi-automated tools for data insertion, integration, documentation, and validation. It is also inadvisable to take on such a project without advice from the ecologists familiar with the data to be incorporated. Our collaboration among ecologists (responsible for experimental design and data analysis), information managers (accountable for data access), and computer scientists (responsible for producing technical solutions) was thus important from the onset of the project.

Here, we present preliminary products and results of the GDI project. We discuss the design, development, and implementation of a centralized database and the tools created to support the integration and importation of these disparate datasets. We also share the lessons that we learned in the integration process, which may be applicable to other ecological and scientific data integration efforts.

## **2. The Grasslands Data Integration (GDI) Project**

Our initial goal was to combine ANPP data from three LTER sites (Jornada Basin [JRN]; Shortgrass Steppe [SGS] and Sevilleta [SEV]), and to include two other grassland sites if possible (Konza Prairie [KNZ] and Kruger National Park ILTER in South Africa [KRU]). The data from each site were collected under different experimental methods under different climate and vegetative conditions, were described using different semantics for experimental units,

---

<sup>1</sup> <http://www.lternet.edu/coreareas/coreintro.html>

<sup>2</sup> [http://daac.ornl.gov/NPP/html\\_docs/npp\\_stat.html](http://daac.ornl.gov/NPP/html_docs/npp_stat.html)

<sup>3</sup> <http://www.fao.org/gtos/NPP.html>

species names, and ecosystem types, and were made available to the project in incompatible syntactic formats. The large number of records from each of the sites led us to explore issues of identifying and fixing data quality problems, and highlighted the need for new tools that would enable both data producers and consumers to explore the data sets in a multi-site database. Exploring data in a multi-site database exposed data quality problems at the site level and raised questions about changes in data collection or reporting over the life of a data set, which we think could lead to improvements in the quality of the data warehoused in the GDI database. In the following sections, we describe activities that took place during the project, the solutions developed to address each problem encountered, and methods used to validate the process and tools that we developed.

### **2.1 Data Collection and Information Management at the Site Level**

Methods for ANPP field data sampling are designed independently by ecologists at each site, and typically change over time. Data are collected by field technicians, sometimes processed or aggregated by the responsible investigators, and then placed into a database local to an LTER site and validated by an information manager. Where observational data are coded by species, a table of species codes and information about the species they represent is also maintained and bundled with the ANPP data.

Previous cross-site integration work by ecologists (e.g., Knapp and Smith, 2001), involved manually acquiring data from each site and combining these data into a new, single-purpose and static database. Even today, most current ANPP data are kept in a local site database and are available upon request in a file format commonly used for exchange between database packages, most often comma-separated-value (CSV) format. The schema of these tables vary greatly among sites and preclude a simple path to automated integration. A lack of semantic metadata regarding the experimental design, how sample replicates should be grouped into statistically-relevant experimental units, and details about how best to aggregate ANPP values can obfuscate the comparison of seemingly-equivalent data between sites. Even species information, which is almost universally gathered, is difficult to integrate because of the use of site-specific codes. There are few processes available for managing species tables, but the accuracy of this information is critical for data analysis.

### **2.2 Cross-Site Data Integration Issues**

In integrating data from multiple sites, we faced various challenges, such as differences in data granularity (whether data were collected by species or growth habit) and differences in site-specific experimental design. In some cases, as at the SGS and KNZ sites, ANPP is measured directly by harvesting total standing crop biomass (Milchunas et al. 1994). At other sites, e.g., SEV, JRN, and Kruger, ANPP is estimated based on species-specific regression relationships between biomass and plant volume or coverage (Muldavin et al. 2008, Huenneke et al. 2002). In still other cases, ANPP is estimated from remotely sensed images and the use of indexes (Paruelo et al. 1997). These different methodologies for collecting ANPP data are conducted at various spatial scales (e.g., one-quarter square meter vs. hectare), and at different temporal (e.g., seasonal or annual) and biological (e.g., species or life form) resolutions. Such differences are common among measures of biotic data.

Each of the sites participating in this project bases ANPP on field measurements, but each has a different number of experimental units at which they collect data. A site might have many plots and each plot many sub-plots; sub-plots might be even further subdivided. We called

the lowest level of sub-plots where data were collected the “experimental unit”. For the integrated database, however, the responsible ecologists emphasized that one should not analyze values at this level but instead aggregate site data at the experimental unit level to a unit appropriate for ecological analysis. We called this level the “sampling unit”, and for our data validation and preliminary analyses averaged ANPP across comparable experimental units, and reported ANPP at various “locations” (sampling units) across sites. Table 1 shows how sampling units and experimental units varied for the sites we worked with. Subsites in the database are distinct because they are considered to have natural differences in ecological characteristics. Some sites perform experiments and certain plots with the same ecological characteristics might be differentiated as a control plot or treatment plot (e.g., burns, livestock grazing).

Site	Sampling Method	Times Measured per year	Years of Data	Number of Vegetation Types or other Relevant Treatments	Number of Sub-Sites	Number of Sampling Units (replicates)	Experimental Units* in each Sampling Unit (plots per rep)	Total Number of Experimental Units (plots)
Kruger National Park (Kruger)	Regression relationships	1	17	35	35	35	9-41	315-1435
Konza Prairie (KNZ)	Biomass harvest	2	5	1	1	2	40	80
Jornada Basin (JRN)	Regression relationships	3	17	5	15	15	49	735
Sevilleta Wildlife Refuge (SEV)	Regression relationships	3	8	3	3	15	16	720
Shortgrass Steppe (SGS)	Biomass harvest	1	23	1	6	3	5	90

**Table 1. Site-Specific methods and number of years of data.** The vegetation types sampled and sampling units at each site were determined by site ecologists; these determine replicates for statistical analysis. The number of experimental units is the number of plots or quads within each sampling unit or replicate.

\* The GDI database, as shown in Figure 1, refers to Experimental Unit as “location”.

Most data within the GDI database have been collected and integrated at the species level, but sites typically use different codes to record species level data. In these cases, observational data are coded by site-specific species codes, and a table or list of those codes and information about the species they represent (the species table) is available. The USDA PLANTS



database is used as the cross-site species table, and we built a general-purpose tool, *Specificik*, to map each site's species codes to the USDA PLANTS codes. Of course, the PLANTS species codes are applicable only to species typically present in the U.S., so adding an international site requires updating the codes database to cover species or plant forms not present in the U.S.

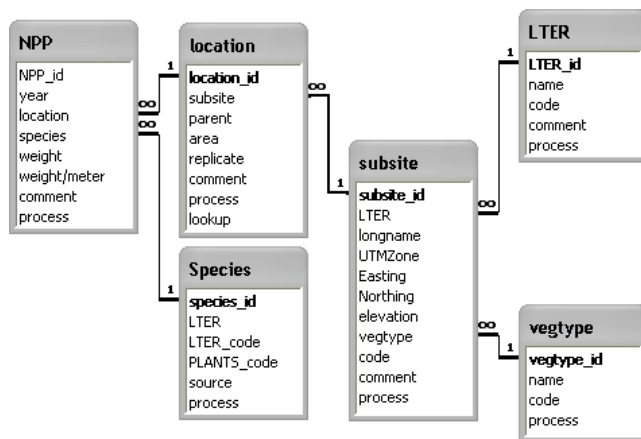
### 3. Methods and Techniques

While constructing a database of one site's ANPP data is relatively straightforward, merging many sites' data at the observation and species levels, and properly combining experimental units into sampling units, is dauntingly difficult and time consuming. As a result, few ecologists to date have analyzed ANPP across sites and those who have typically have limited the granularity of their analysis, leaving many potentially important variables among sites unexplored. For example, the EcoTrends database aggregates ANPP data into total annual values per site, vegetation type or treatment and does not differentiate by species.

We identified three key areas for the GDI project: schema design, a robust process for integrating data, and species integration. The first step was to develop a data model that represents the complexity of ANPP data across sites and that is simple to explain and use; the second to write scripts to process each site's dataset as this processing had to be repeated several times – to correct data errors in the original datasets and to add additional years of data as they became available. These scripts will be used in the future as data for additional years and additional sites become available. The third step was to develop a program to map site-specific species codes to PLANTS codes.

#### 3.1 The GDI Data Model

Because of data quality issues (primarily data type and referential integrity errors) identified during the first data loads and integration, we determined that the best design would be a single, centralized database to be updated periodically, rather than a virtual database or index from each LTER site's online ANPP database.



We chose to design the ANPP integrated database so that the NPP table is its primary focus. Each row of that table contains information on how much NPP was collected per year, for what species (or growth habit), and where and when it was collected. Figure 1 shows this schema, as generated by MS Access. Each table name appears in grey highlight above a list of attributes (rows) for the table; the cardinality of all relationships between tables is many ( $\infty$ ) to 1, as shown by solid lines connecting tables. All tables contain a unique identifier (primary key), which is first in the list of

attributes and is the name of the table followed by "\_id". Many tables also include a comment field, which at this point is free text, and a process field, which is a set of numbers corresponding to processes listed in a metadata document, so that it is possible to recreate exactly the queries and scripts that affected each row to process the data.

A record in the **NPP table** contains the following attributes and relationships to other tables: a unique identifier NPP\_id, Year when NPP was measured, location of the measurement(s) as a foreign key relating to the location table, code for species or plant type as a foreign key relating to the Species table, weight of NPP as measured (units differ per site), and weight of NPP adjusted to grams per meter squared. Thus, from this table, a researcher can determine observed ANPP (in a common unit, grams of biomass per meter squared) within a given location (an area defined on a per-site basis, referred to as the sampling unit) over one calendar year. While growing season is used by many sites to report NPP, we did not report data at this level of time granularity since we could not compare growing season across sites as they differ.

The **Location table** contains information about the experimental unit to which the measured NPP value is associated. In particular: subsite relates to an area of ecological interest for which biome and climate data are available in the Subsite table, area is the size of the (plot) location in meters squared, replicate denotes membership in one of the sampling units within the subsite; each sampling unit contains 5-50 plot-locations. lookup is used for validation during the integration process; it is not meaningful to the ecologist. parent is not currently used; before clumping many experimental units into one subsite, location was organized into a hierarchy of plots, subplots, subsubplots, etc.

The **Subsite Table** describes the subsite to which (plot) locations belong. It identifies the [LTER] research site at which the measurements were taken, and is typically identified at the research site by a name and code. The Subsite table also contains geographic coordinates and information (UTMZone, Easting, Northing, and elevation) as well as vegtype (vegetation type or treatment of interest) for comparative analysis within the larger LTER research site. vegtype (vegetation type) is a coded value described in the Vegtype table, which contains a full name and code for the vegetation type. Sampling units (replicates) and experimental units (location) are further explained in Section 3.2, where we articulate the data transformation process from site-specific data to the integrated database.

The **Species table** relates the standardized USDA PLANTS species code in the NPP table to a site-specific species code. Each PLANTS species code could correspond to multiple LTER site-specific codes at different sites, and (over time) any one LTER code could relate to more than one PLANTS code.

The major challenge we identified in loading data in the GDI schema was cleaning up referential integrity for species codes and deciding which USDA PLANTS code to associate with each site-specific code. That conversion process prompted us to create a tool which may be useful for broader work with species-coded data in the botanical domain. This tool is explained in the Section 3.3.

### **3.2 Data Transformation and Integration**

Prior to adding each site's data to the integrated database, we carried out two steps: *transformation* to the NPP observation table format and *integration* with data from other sites. The separation of these steps allowed easy re-integration of data when changes occurred in the experimental design, as location, sub-site and species code tables were loaded only once. Each LTER provides biomass data, either calculated or directly measured. If it is not in a yearly format, seasonal data is combined into years. These transformations are handled with ad-hoc scripts specific to each site's data. Future data submissions will be required in the "observation" format and validated and processed with a common script.

Once each dataset is formatted as a series of NPP observations, it must be integrated into the central database with enough contextual data to allow for meaningful statistical comparison across research sites by *location* or by *species*. Individual plot areas range from a quarter of a meter squared to two hundred meter line transects, so plots are not directly comparable. Each individual plot (location) is assigned to a statistical sampling unit, which is an aggregation of co-located plots with similar soil and vegetation types to allow statistically meaningful analysis. Individual sampling units are designated by the ecologists as containing enough data to be statistically meaningful and contain between five and fifty plots. In the GDI database, we call the sampling unit the replicate. In addition, each plot (location) is assigned to a subsite within the LTER – an ecologically meaningful geographic designation for which biome, geographic location, and climate data are available. This contextual data allows aggregation of data (beyond species or plant code) for analysis at three additional levels: research site (LTER), subsite, or sampling unit. Mechanical parsing of most species information proved successful, as species information is generally provided in CSV format, though occasional human intervention was required where a site's species table was not syntactically self-consistent.

### 3.3 Species Code Conversion

Site-specific species tables are typically not congruent with USDA PLANTS codes since most sites use site-specific codes. The USDA PLANTS codes, on the other hand, use rules defined by the International Code of Botanical Nomenclature for assigning botanical names. Although some taxonomists have defined new standards for botanical naming in line with information science best-practices<sup>4</sup>, binomial names remain the dominant standard and are the only system of plant taxonomy that is generally-accepted. Complicating integration, even binomial names used correctly today may become inaccurate over time. Recent years have seen trends towards plant reclassification based on DNA evidence, making some names obsolete. If it is discovered that a species has been incorrectly assigned to a genus then the genus given as part of its binomial name would lead to inaccurate analysis between genera.

The USDA PLANTS database contains placeholder markers for those obsolete names, and encodes information about which currently-accepted name is synonymous. We decided that the best practice would be to use the USDA PLANTS species code in our database, and create a table that provides the correspondence of that code to the name/code used by the LTER site, and thus to make use of the USDA PLANTS synonym infrastructure. The USDA PLANTS codes have been established as the standard for identification of plants species for the four U.S. LTER sites<sup>5</sup> in the GDI database, and the correct USDA PLANTS code must be determined to avoid falsifying species coded data. Some information of interest for many ecologists, such as carbon pathway, seems to be absent from USDA PLANTS, and must be maintained externally. This presents problems as we found no single authoritative source for such information. However, some information not kept by many sites, but useful in analysis and reporting such as threatened, endangered and invasive status and common name, is available from the USDA PLANTS database.

---

<sup>4</sup> See Phylocode (<http://www.ohiou.edu/phylocode/index.html> and <http://www.ohiou.edu/phylocode/PhyloCode4b.pdf>) and Biocode <http://www.bgbm.org/iapt/biocode/>.

<sup>5</sup> The non-U.S. iLTER site in our database, Kruger, did not record ANPP by species, so the fact that the USDA PLANTS codes do not cover South Africa was not an issue. If we include other ANPP datasets for non-U.S. sites that provide ANPP by species, we will need a species table similar to PLANTS for those sites.

Just as data errors are common, spelling mistakes are frequent enough in binomial names for species at LTER sites to have required a manual process at many points in the conversion. A manual (or at least interactive) process is likewise highly desirable as it brings species errors to the attention of the information manager and the ecologist. The responsibility to make a determination of correct equivalent USDA PLANTS code in such situations is not a technical decision, and will rest with the contributing site. To facilitate this process we have developed a web-based application, *Specifik*, described in Section 4.2.

#### **4. Results and Discussion**

The GDI database, as of March 20, 2008, contained 113,500 distinct NPP observations from five sites, and was 73 MB in size. Because the database creation process brought data errors to light, the finished database contained fewer errors than the source data. Normalization highlighted errors of absence: data missing from certain plots or certain years and blank species or mass data. Integration highlighted errors of context: species with entries in the data but not the species tables, data from mislabeled or nonexistent plots, or plots with bad coding information. In addition, some basic validation checks removed observations with negative or zero weight, and observations for years outside the known span of the experiments. Questionable data were resolved by conversations with data providers. As we go to press for this paper, the GDI database is not yet released for general download, but it can be requested from the Sevilleta LTER Information Manager.

##### **4.1. Preliminary Validation**

The database allows comparison of ANPP between LTER sites and vegetation types, and we found that the discipline of data integration and preliminary scientific analysis can lead to improving data quality for better subsequent analysis. For example, an early statistical comparison of three LTER sites, JRN, SGS and SEV erroneously suggested that Jornada was significantly more productive than similar grassland sites despite the fact that it is the warmest and driest. This led the Jornada site to update its regressions and helped to emphasize the fact that the prevalence of a single species (*Yucca elata*) at one site can influence cross site analysis.

##### **4.2 Specifik**

*Specifik* is a web-application that we wrote to assist a user in adding USDA PLANTS species codes to a species table. Given a CSV-formatted table of species information, it asks the user a few simple questions to determine the dataset's taxonomic ontology. It then asks users to select a USDA PLANTS code for each species from a list of likely matches. If a user is unsure which alternative is correct, the tool allows the user to defer the assignment, and to provide a manually processed code at a later time. Once codes have been selected for every species, users are given a copy of their species code table with USDA PLANTS codes added. We hope that by contributing an easy-to-use tool to facilitate this process, more sites will maintain USDA PLANTS codes in their own databases, facilitating future work by us and others who hope to integrate data and analyze species-coded data. *Specifik* defers to the USDA PLANTS database on issues of taxonomy, as there are taxonomists and biologists who ensure that the species information therein is current and correct. The USDA PLANTS database also contains most species metadata documented by each LTER site, such as genus, author, family and form, but

that information may be incomplete for many species. *Specificik* is open source and freely available for download from the internet.<sup>6</sup>

## 5. Lessons Learned, Conclusions, and Future Work

Perhaps the most critical lesson learned from this project is the fact that ANPP is but one of a class of ecological data dubbed *response variables*, e.g., measurements of primary productivity (NPP, biomass, cover, etc.) or diversity (species richness, species diversity, community dynamics, etc.). While useful in and of themselves, ecological response variables are significantly more useful if environmental drivers, or context variables, are available in such a way that correlations can be drawn between environmental drivers and responses. To model change in ANPP over time, contextual data is required. Climate is a key driver of year-to-year changes in production, so models of change in ANPP over time would include climate data such as Palmer Drought Severity Index (PDSI), precipitation (e.g., growing season vs. non-growing season, totals, extreme events), temperature data (e.g., growing season vs. non-growing season, minimums, maximums). Other important contextual data might include landform for each location (e.g., slope, aspect, soils, and elevation), and land history (e.g., grazing and fire). Soil moisture, and atmospheric and soil chemistry could also impact ANPP. Links between a GDI database website and data stores such as EcoTrends,<sup>7</sup> ClimDB/HydroDB,<sup>8</sup> and the National Atmospheric Deposition Program<sup>9</sup> could facilitate analyses of queried NPP data with contextual data.

Response variables (aka biotic data) are significantly more complex than environmental drivers such as precipitation or temperature which have a history of data recording and reporting, but both are needed for answering important ecological questions (Peters et al, 2008). Secondly, we discovered that the process of putting complex data into a database can go a long way to improving data quality; analysis of cross-site data integration products provides additional quality control. Of course, the additional effort required to standardize codes (e.g., for species) and semantics (e.g., sampling units or plot granularity, vegetation type) is significant. Thirdly, interdisciplinary collaboration and teamwork during the data integration design process are key to success. Without any one of our stable three-prong foundation (ecologists, information managers, and computer scientists), this project could not have succeeded. The challenges to make the GDI successful and useful required a team approach with communication among ecologists, information managers and computer scientists. Fourthly, we assert that once a general schema has been developed and well tested, and easy to use tools and processes established, the burden of uploading data to a data repository, such as the GDI, should lay with the site rather than with a central curator. While a curator must take ultimate responsibility for including a certain site's data in a repository, only the local information manager has adequate in-depth understanding of the data to perform the required data transformations.

In conclusion, we suggest that the GDI project has created a sustainable, streamlined system for transforming, integrating, validating and analyzing ANPP data. It also required time and attention to identification of data quality issues during data transformation and integration, and a fairly deep understanding of how the data would eventually be analyzed. This collaboration also resulted in a species code standardization tool that can be used for other

---

<sup>6</sup> <http://alala.evergreen.edu/~mallettj/specifik/>

<sup>7</sup> <http://www.ecotrends.info>

<sup>8</sup> <http://www.fsl.orst.edu/climdb>

<sup>9</sup> <http://nadp.sws.uiuc.edu/>

synthetic research projects. The GDI project has helped information managers become more aware of LTER synthesis projects, such as EcoTrends, as ANPP represents a core area of research for the LTER Network and a long-term ANPP dataset would serve as a foundation for cross-site and synthetic research future scientific endeavors (Baker, BioScience). Standardized, structured, centralized and accessible repositories of such data facilitate the work of both information managers and ecologists.

### **Acknowledgements**

NSF Canopy Database Project (NSF Grants: DBI-0417311, DBI-0319309), JRN-LTER (NSF Grant: DEB-0080412), KNZ-LTER (NSF Grant: DEB-0218210), SEV-LTER (NSF Grant: DEB-0080529), and SGS-LTER (NSF Grant: DEB-0217631). We also thank Alan Knapp, Debra Peters, Mark Servilla, Susan Stafford, Bob Waide, and the LTER and Kruger National Park field crew members and technicians, for helpful suggestions or work on this project.

### **References**

- Baker, K.S., B.J. Benson, D.L. Henshaw, D. Blodgett, J.H. Porter, and S.G. Stafford. 2000. Evolution of a multisite network information system: the LTER information management paradigm. *BioScience*, 50:11, pp 963-978.
- Huenneke, L.F., J.P. Anderson, M. Remmenga & W.H. Schlesinger. 2002. Desertification alters patterns of aboveground net primary production in Chihuahuan ecosystems. *Global Change Biology* 8:247-264.
- Knapp, A.K. and M.D. Smith. 2001. Variation among biomes in temporal dynamics of aboveground primary production. *Science* 291:481-484.
- Knapp, A.K., M.D. Smith, S.L. Collins, N. Zambatis, M. Peel, S. Emery, J. Wojdak, M.C. Horner-Devine; H. Biggs, J. Kruger, S.J. Andelman. Generality in Ecology: Testing North American Grassland Rules in South African Savannas. *Frontiers in Ecology and the Environment* 2(9):483-491.
- Milchunas, D.G., J.R. Forwood & W.K. Lauenroth. 1994. Productivity of long-term grazing treatments in response to seasonal precipitation. *J. Range Manage.* 47:133-139.
- Muldavin, E.H., · D.I. Moore, S.L. Collins, K.R. Wetherill & D.C. Lightfoot. 2008. Aboveground net primary production dynamics in a northern Chihuahuan Desert ecosystem. *Oecologia* (155):123-132
- Parton, W.J., J.M.O. Scurlock, D.S. Ojima, D.S. Schimel, D.O. Hall, M.B. Coughenour, E. Garcia Moya, T.G. Gilmanov, A. Kamnalrut, J.I. Kinyamario, T. Kirchner, S.P. Long, H-C. Menaut, O.E. Sala, R.J. Scholes, and J.A. van Veen . 1995 . Impact of climate change on grassland production and soil carbon worldwide. *Global Change Biology*. 1: 13-22
- Peters, D.P.C., Guest Editorial, *Frontiers in Ecology and the Environment*, Special Issue on Connectivity and Continental Scale Research, June 2008.
- Paruelo, J.M., Epstein, H.E., Lauenroth, W.K., & I. C. Burke. 1997. ANPP Estimates from NDVI for the Central Grassland Region of the United States. *Ecology* 78(3): 953-958.
- USDA, NRCS. 2008. The PLANTS Database (<http://plants.usda.gov>, 29 April 2008). National Plant Data Center, Baton Rouge, LA 70874-4490 USA

## **BUILDING A “CYBER FOREST” IN COMPLEX TERRAIN AT THE ANDREWS EXPERIMENTAL FOREST**

**Donald L. Henshaw<sup>1</sup>, Fred Bierlmaier<sup>2</sup>, Barbara J. Bond<sup>2</sup>, and Kari B. O’Connell<sup>2</sup>**

<sup>1</sup>U.S. Forest Service, Pacific Northwest Research Station, Corvallis, OR 97331, USA

<sup>2</sup>Department of Forest Science, Oregon State University, Corvallis, OR 97331, USA

### **Abstract**

Our vision for a future “cyber forest” at the Andrews Experimental Forest foresees high performance wireless communications enhancing connectivity among remote field research locations, station headquarters, and beyond to the university and outside world. New sensor technologies and collaboration tools foretell exponential increases in data and information flow to accommodate both research and education. We envision improving data transmission speed and bandwidth, enveloping the Andrews in a wireless cloud, installing new technologies for near real-time access to sensor networks, and assuring quality and management of streaming data. The remote location of the Andrews Forest, far beyond the reach of conventional communication infrastructure, coupled with steep mountainous topography and very tall trees present significant challenges to the realization of this vision. This paper explores the innovative approaches being tested to provide real-time access to important data streams and unique educational opportunities.

**Keywords:** cyberinfrastructure, wireless networks, sensor arrays, complex terrain, Andrews Experimental Forest, Long-Term Ecological Research (LTER)

### **1. Introduction**

New technological capabilities in environmental sensing and communications are revolutionizing the science questions asked and study experiments performed to reveal previously unobservable phenomena (Estrin et al. 2003). Sensor networks are enhancing the scales at which scientists perceive the natural world, increasing both spatial intensiveness and extensiveness and temporal frequency (Porter et al. 2005). Communication technologies are similarly expanding the education and outreach capacity of field stations allowing distance education and virtual field trips through real-time video and data connections with classrooms and meeting rooms around the world.

Recognizing the potential of these new technologies, field stations across the United States are beginning to “go cyber”. However, strategies that work at one location may not be simply transferred to all others due to local constraints and circumstances. At the Andrews Forest, for example, we face extreme challenges due to the rugged terrain, massive trees, remote location, and large area of the experimental forest. We are developing a comprehensive plan to build a “cyber forest” at the Andrews that meets these challenges. Our plan includes increased bandwidth to the home institution, improved field-to-headquarters data transmission, complete wireless coverage of the Andrews Forest, and dynamic data management and quality control tools to accommodate streaming data from sensor arrays in near real-time (*Figure 1*). This paper will primarily discuss improving bandwidth, development of the “wireless cloud” and field-to-headquarters communications, and our approach to examine the feasibility for establishment and implementation. This extended wireless local area network (WLAN) will provide a quantum

leap in capability for both research and education at the Andrews Forest as well as provide a prototype that might be applied in other mountainous research areas. We anticipate researchers posing new science questions, such as understanding the influence of complex terrain on ecosystem structure and function, where topography creates enormous environmental variability that demands establishment and real-time access to sensor arrays.

Wireless communication systems (telemetry) were first installed in 1994 in a few of the high elevation meteorological and stream gaging stations at the Andrews Forest. Near real-time web displays of this data proved useful in assuring successful operation of remote sites and provided data access even when physical access was restricted due to heavy snow or debris avalanches. The telemetry network has since been expanded to include a more extensive hydrometeorological measurement system, and an automated system allows near real-time Internet access and graphical displays of data. This original Campbell Scientific radio telemetry system using VHF radios at a licensed frequency of 151.65 MHz is no longer supported by Campbell Scientific, requiring future replacement of the radio modems and associated repeaters and base station.

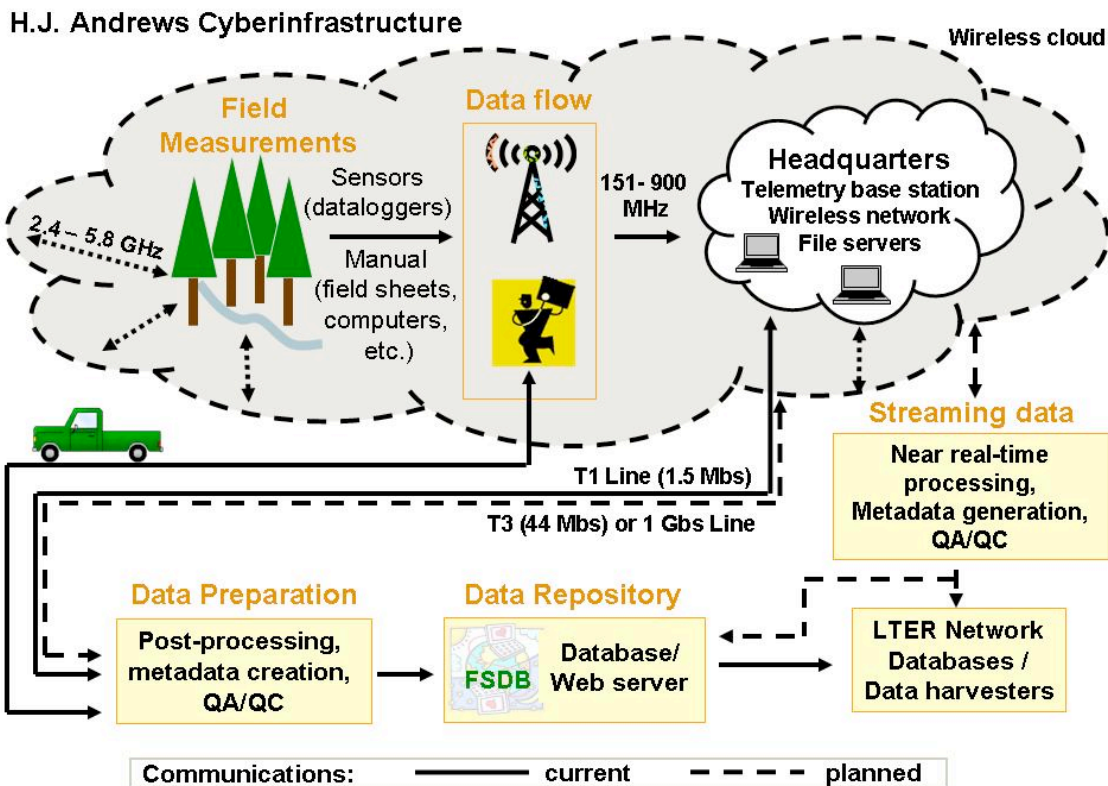


Figure 1. Current and planned infrastructure to support the Andrews “cyber forest”. A wireless cloud (WLAN) blanketing the entire Andrews Forest is envisioned allowing transmission of sensor network data and Internet access throughout the mountainous topography. Improved bandwidth and communication rate are planned. Applications for near real-time data processing, data validation and metadata generation are anticipated.

Communication from the Andrews headquarters to Oregon State University (OSU), a distance of about 160 km, was limited to 56 Kbs until establishment of a T1 line (1.5Mbs) in 2001 using existing telephone cables. Currently, the headquarters site hosts a wireless local area network (802.11 standard) composed of a series of access points and wireless bridges that allow connection throughout the headquarters site including office and apartment buildings. However,



voice communications throughout the forest are limited to handheld radios that do not work in many locations.

## 2. Methods

Our methods involve examining the science needs for these enhanced communications, explore the tremendous challenges to building the cyber forest including major issues of topography and bandwidth, and begin testing new technologies in forested conditions.

*Science needs:* Rapid increases in digital data collections and sensor array deployment demand improved data transmission and expose limitations in available bandwidth and performance, especially when considering the following current uses and future needs:

- Streaming sensor array data to the Internet at fine temporal resolution - atmospheric studies, hydrometeorological measurements, and snow and subsurface hydrology studies are underway with others planned.
- Planned deployment of acoustic sensors for biodiversity studies.
- Collaboration technologies including videoteleconferencing with multipoint software, voice over IP and video transmission from the field.
- Server backups and synchronization of mirrored servers at the university.
- Web cams and frequent transmittal of large-size digital images are in operation (i.e., university web cam sends images every 30 seconds), and web cams are planned for bird/animal tracking and phenology monitoring.
- Data intensive applications, e.g., GIS, visualization

*Topography:* The topography and vegetation of the Andrews Forest have created the need for more extensive sensor arrays while also presenting a significant obstacle in the transmission of collected data. The forest is broadly representative of the rugged mountainous landscape of the Pacific Northwest with elevation ranges from 410 m to 1630 m and old-growth forest stands among the tallest and most productive in the world. The richness ascribed to topographical-position has demanded that sensor arrays be established with high station density and wide variations in topography and canopy cover (Daly et al. 2007). The terrain and canopy cover seriously limit transmission range and bandwidth, but the Andrews Forest offers a unique research challenge in developing wireless connectivity in remote mountainous landscapes.

*Approach:* The following approaches were used to explore potential improvements in data transmission speed and bandwidth, new technologies for near real-time access to sensor networks, and management and quality assurance of streaming data:

- Explore costs and feasibility of upgrading the T1 (phone line) transmission rate from the Andrews Forest to OSU including a) fiber-optic communications at 1Gbs and b) building wireless links over 160 km of mountainous terrain
- Explore costs and feasibility of blanketing the entire site with a wireless communications network by a) examining existing technologies, b) using line-of-site (GIS viewshed) software to determine an optimal wireless bridge configuration, and c) using point-to-point link estimator software to determine maximum rate of throughput given topography and vegetation obstruction

- Deploy an extensive sensor array in a small watershed near the headquarters designated as a “cyber watershed” to monitor ecohydrological processes while exploring new technologies and testing the efficiency of spread spectrum radios through dense vegetation
- Prototype a new data model to provide efficient throughput and near real-time quality assurance checking of data streams from multiple sensor networks

### 3. Results

*Bandwidth:* Dark (unused) optical fiber which could provide gigabit performance from the site to OSU is simply not available for the final 50 km to the site. Construction using existing aerial paths (power poles) for 40 km is estimated close to \$1M, and construction over the final 10 km to the Andrews is more expensive with trenching required through rocky material and estimated cost is \$2M. Given the high cost for an optical fiber cable connection, other options are considered. Options include building wireless links over 160 km of foothills using ridge-top towers to directly connect with OSU, or alternatively building a wireless connection to the end of the dark fiber. In both cases, solar-powered towers are necessary but problematic in winter because of low available light and difficult access due to snow.

The critical element for alternative wireless pathways to OSU as well as establishing site-wide wireless communications is the link from the valley-bottom headquarters to a ridge-top tower (Quarry), which is challenged by an intervening ridge and tall vegetation. A more expensive, non-line-of-site wireless bridge with maximum throughput of 300 Mbs is considered for this segment. Point-to-point link estimator software was used to conduct a preliminary path analysis and a theoretical mean throughput of only 92 Mbs is estimated due to free-space path loss and obstructions to line-of-site. Given these difficulties in providing a wireless link to the Quarry ridge site, trenching from the headquarters to the ridge to lay power and optical fiber cables becomes a viable option and costs are being estimated. Laser technology is also considered but power and direct line of sight requirements coupled with problems due to mist and fog conditions are serious impediments. For any option we will maintain the T1 connection as an alternate pathway to assure continual communication. In the interim it is likely that a strategy for sharing bandwidth (e.g., nighttime server backup or scheduled data downloads) between the headquarters and OSU will be needed to prevent transmission bottlenecks.

*Wireless cloud:* We envision blanketing as much of the Andrews Forest as possible with a wireless communications network, prioritizing roads and areas with intensive study sites. Intensive study sites targeted will include climate or stream gaging stations, a small “cyber watershed” in the SE corner (WS 1), and a transect along an elevation gradient designated for new biodiversity and climate change research.

To establish complete wireless coverage of the Andrews Forest, a series of wireless Ethernet bridges starting at the headquarters will be established on communication towers. Each of these bridges will be on high points or ridges with excellent visibility to a number of other ridge sites and a clear unobstructed view to a large forest area. At each of these bridge points, multiple wireless access points will be daisy-chained and established with directional antennas. Directional antennas (as opposed to omni-directional) will yield more power per area to target intensive study areas and provide optimal coverage within the forest boundary. Researchers will be able to connect to the network with WiFi-enabled laptop computers while in close proximity to an access point, but will likely require higher power client devices with integral directional antennas to connect throughout much of the Andrews Forest. While use of the 802.11 standard

for the wireless access points is planned, the WiMax or 802.16 standard is being considered for its non-line-of-sight capability. Unlicensed frequencies of 5.8 GHz and 2.4 GHz for the wireless bridges and access points, respectively, are planned; however, 900 MHz radios might be necessary for access points with better penetration through tree canopies gained at the expense of bandwidth. New 900MHz wireless modems will take advantage of the Ethernet bridge infrastructure. A single wireless modem connected to the bridge at each tower site will be able to consecutively transmit data from several dataloggers at scheduled intervals. Faster transmission rates will allow for near real-time access.

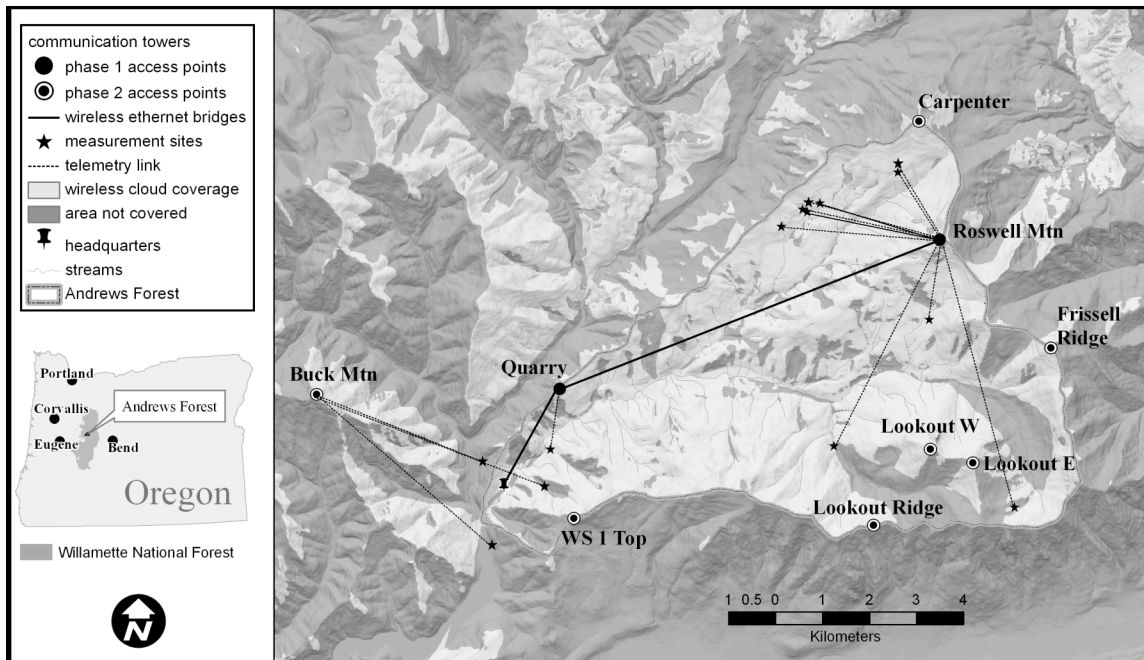


Figure 2. Phase 1 access points and two wireless bridge links should provide 80% WiFi coverage to the Andrews Forest as shown. Coverage is indicated by a light color; dark colored areas are predicted to be in a “wireless shadow.” Phase 2 access points and point-to-multipoint wireless bridges will complete the “wireless cloud” with >95% coverage (estimated from line-of-sight analyses, ArcGIS Toolbox/surface raster/viewshed).

A line-of-sight analysis indicates that a total of 10 communication tower sites linked with wireless bridges will be necessary for complete (>95%) wireless coverage of the site (Figure 2). Interestingly, only three tower sites are needed to provide 80% coverage, including the 11 current telemetry sites, and installation of these are planned as a first phase. Phase I installation will include two key point-to-point bridges between the headquarters and a high-elevation site (Quarry) and from this site to a distant ridge (Roswell Mountain). The Roswell site will provide line-of-sight coverage to the high-elevation watersheds and climate stations. Phase II will install more wireless bridges configured as point-to-multipoint to allow multiple pathways for communication.

*Test deployment:* An extensive sensor array is installed in a small watershed near the headquarters designated as a cyber watershed to monitor ecohydrological processes, and is accompanied by a telemetry system that uses modern 900 MHz spread spectrum wireless modems to transmit data from multiple dataloggers. The topography forces the use of a two-radio repeater on a nearby ridge top, but high-frequency data is reliably transmitted to headquarters even when radios are installed beneath heavy canopy cover. In conjunction with

the cyber watershed concept, researchers have been developing new sensor technologies on the Andrews Forest including distributed temperature sensor (dts) data collection along 4 km of fiber optic cable in two watersheds (Selker et al. 2006) and low power sensor network development employing RF harvesting (Le et al. 2006) .

Time Cost	Years				
	0	1	3	5	10
0	Bandwidth sharing				
\$K		Data model for streaming data		Replace Met/Hydro VHF radio telemetry w/ 900MHz wireless	
50					
100	Cyber watershed		Phase I wireless bridges 92 Mbs	Phase I data access points 80% coverage	Phase II data access points 95% coverage
300				OSU 300 Mbs wireless/fiber; Fiber/power to Quarry Ridge	Phase II wireless bridges
1000+					OSU 1Gbs Fiber connection

Figure 3. The general timeline for cyberinfrastructure establishment will proceed based on funding resources and success in establishing connections in complex terrain.

*Data model:* The arrival of new sensor arrays has resulted in an exponential increase in the quantity of data collected and places significant strain on existing resources used to quality assure and archive data. Current metadata-driven, post-collection processing and data validation methods struggle to keep pace with timely archiving of sensor data. Resolving many of these issues of data stream management is a community problem and the Andrews information managers will track progress and solutions emerging from other Long-Term Ecological Research sites and observatory networks, but at the same time we continue to push for local solutions. A short description of one current effort follows.

A new data model and processing system is being established to manage and provide for seamless acquisition of hydrometeorological data streams. Features include:

- Master catalog of all sensor arrays (or data loggers) with operational date ranges
- Detailed table of all sensor array (or data logger) configurations including measurement metadata and ordered lists of measured parameters
- Normalized data tables of all sensor data with qualifier codes to preserve all raw data streams with appropriate metadata for later resampling
- Automatic screening of data against prescribed data limits and assignment of qualifier codes for each measurement value
- Data acquisition and graphical interface to monitor data streams and provide comparative graphs of questionable data.

For other sensor networks, the headquarters base station will facilitate direct streaming of data to campus labs where processing protocols are established.

#### 4. Conclusions

Multiple challenges face the scientists and educators at the Andrews Forest before the vision of a “cyber forest” is reached. New cyberinfrastructure capability is necessary to assure

near real-time data access, storage and backup requirements, efficiency of processing, and data quality. The remoteness and distance from conventional communication pathways stand in the way of the desired high performance connectivity between the site and university. The forest topography and tree canopy height pose unique challenges compared with more typical telecommunication installations. An approach to the development of an all encompassing WLAN is established, and the planning and installation of new technologies and sensor arrays proceeds.

While relying on expert advice to guide our planning, it is clear that there are many uncertainties in developing this telecommunications network. Resources for reliable fiber optic solutions may not be available, and knowledge of whether planned wireless bridges will work effectively until tested in complex terrain is unknown. The general implementation timeline (*Figure 3*) will have to be modified if key wireless segments prove untenable. This is a new frontier in ecological studies to conduct such research in complex terrain, and it will require flexible timelines and significant resources to build necessary bandwidth for rapid communications. Once achieved, however, our vision for cyberinfrastructure will provide unique new opportunities for science and education.

### **Acknowledgements**

The authors would like to acknowledge Theresa Valentine for the GIS analysis and map and Mark Klopsch for technical advice. Funding for this work was provided by the NSF Andrews LTER grant (DEB0218088) and the U.S. Forest Service PNW Research Station.

### **References**

- Daly, C., Smith, J. I., and McKane, R. 2007. High-resolution spatial modeling of daily weather elements for a catchment in the Oregon Cascade Mountains, United States. *Journal of Applied Meteorology and Climatology*, pp. 1565-1586.
- Estrin, D., Michener, W. and Bonito, G., 2003. Environmental cyberinfrastructure needs for distributed sensor networks: A report from a National Science Foundation sponsored workshop. NSF Workshop 12-14 August 2003. Scripps Institute of Oceanography, La Jolla, California.
- Le, T., Mayaram, K. and Fiez, T. S., 2006. Efficient Far-Field Radio Frequency Power Conversion System for Passively Powered Sensor Networks. *Proceedings of the September 2006 IEEE Custom Integrated Circuits Conference*, pp. 293-297.
- Porter, J., Arzberger, P., Braun, H.-W., Bryant, P., Gage, S., Hansen, T., Hanson, P., Lin, C.-C., Lin, F.-P., Kratz, T., Michener, W., Shapiro, S. and Williams, T., 2005. Wireless sensor networks for ecology. *BioScience* 55:561-572.
- Selker, J.S., Thévenaz, L., Huwald, H., Mallet, A., Luxemburg, W., van de Giesen, N., Stejskal, M., Zeman, J., Westhoff, M. and Parlange, M. B., 2006. Distributed Fiber Optic Temperature Sensing for Hydrologic Systems. *Water Resources Research*. doi: 10.1029/2006WR005326.

## RANGE CHECKS OF COASTAL ENVIRONMENTAL MONITORING DATA

Lei Hu<sup>1</sup> and Brenda Leroux Babin<sup>2</sup>

<sup>1</sup> Dauphin Island Sea Lab, Dauphin Island, Alabama. <sup>2</sup> Louisiana Universities Marine Consortium, Chauvin, LA

### Abstract

Data Quality Control (QC) is an important component of Ocean Observation Systems. Data quality control ensures that good and valid information is passed to researchers, educators, and the public for planning and decision making. Range checking is an initial key step in Data Quality Control. It checks the data, and flags it for further investigation if necessary. This paper explains the planning, programming, and implementation of range checks on monitoring data, at Dauphin Island Sea Lab and Louisiana Universities Marine Consortium (LUMCON).

**Key words:** Data quality control, range check

### 1. Introduction

The purpose of a real-time monitoring system is to provide researchers, educators, and the public with quality data for planning and decision making. Both Quality Assurance and Quality Control are important to ensuring good quality data. During the first stage, Quality Assurance involves all of the actions taken while data is being collected to ensure that the data are "good." In the second stage, Quality Control involves the steps taken after the data is collected to ensure the data is "good" before the data is released to the public. At the first meeting of the Quality Assurance of Real-time Oceanography Data (QARTOD I), several steps were identified for the second stage, Quality Control. The first step is simply to insure completeness of the data set. This involves several components, such as: a transmission check, making sure a valid message was received (e.g. checksum validation); a time stamp validity check; and a check for service schedule of hardware. The second step is to perform various range checks on the data, including initial gross range checks based on instrument specifications and range checks based on the climatology of the area. Finally, additional quality control measures include checks against trends in the data and checks against nearby stations (NOAA 2004). QARTOD I further suggests that whenever possible this quality control should be automated and yet contain a manual check by a human component within a few days of disseminating the data.

Following QARTOD I, the data managers from Dauphin Island Sea Lab and Louisiana Universities Marine Consortium met to discuss possible ways for implementing a quality-control plan for each system. Because both systems already had some form of data-completeness check in place, it was decided that we would start with an automated range-check system. Several conditions were critical to implementing this range checking system. The system had to be flexible enough to allow easily for the addition of sensors, and the system had to allow for different ranges for each site.

### 2. Planning

Automation of range checking of environmental monitoring data was implemented in September 2006 at the Dauphin Island Sea Lab, based on ranges established by NOAA's

National Estuarine Research Reserve (NERR). The range of measurement specifications for both meteorological data and water quality data are listed in Table 1 (Small 2004a, b). Although LUMCON is not part of the NERR program, they decided to establish the same range checks to maintain consistency among the systems and to share in the implementation.

Table 1: Range Checking Criteria based on ranges established by NERR.

Meteorological Data	
Air Temperature	-10 °C < And < 50 °C
Wind Direction	0 degree < And <360 degrees
Wind Speed	0.5 m/s < And <30 m/s
Barometric Pressure	900 mb < And < 1060 mb
Solar Radiation	0 KW/m2 < And < 1.5 KW/m2
Quantum Radiation	0 μE/m2/s < And < 2700 μE/m2/s
Precipitation	0 mm/min < And <3 mm/min
Relative Humidity	0% < And < 100%
Hydrographic Data	
Water Temperature	-5 °C < And < 45 °C
Water Level	At Dauphin Island Station: -3.0 ft. < And < 5.0 ft. At Meaheer Park Station: 1.0 ft. < And < 6.0 ft. At Middle Bay Light Station: 2.0 meters< And <6.0 meters
Salinity	0 ppt < And < 40 ppt
Dissolved Oxygen Percent:	0% < And < 200%
Dissolved Oxygen mg/L	0 mg/L < And < 20 mg/L

### 3. Programming

The first step was to create the flag fields that stored the values for flags (see Table 2 below). For each parameter, one flag field was created in the same table. The range checking logic was translated into SQL (Structured Query Language) Server stored procedures (Appendix 1). The stored procedures run twice every hour, fifteen minutes after the data are harvested. Data checked are the one minute meteorological data, and the thirty minute hydrographic data. The quality of real-time data is described by an aggregate quality flag recommended by QARTOD I:

- 9 = missing value
- 0 = quality not evaluated
- 1 = bad
- 2 = questionable/suspect
- 3 = good

This stored procedure flags the data as -9, 2, or 3, because we think human checks should be performed in order to determine if the flag should be 1. Criteria are hardcoded in the stored procedure logic. In the two years since the range check has been implemented, the criteria were updated only once. Criteria and flag are both defined on a per-parameter basis (Table 3), no flags are assigned to the record as a whole. The stored procedure uses an air-temperature flag as a check point to see if a record has been checked. If the air temperature flag has no value, then all parameters of that record need to be checked and flagged.

Table 2: Meteorological data table before range check

Year	Day	Minute	Watertem p	Watertem p flag	Solarrad	solarrad flag	quantumra d	quantumra d flag
2008	178	800	27.68	Null	0.257	Null	647.6	Null
2008	178	801	27.68	Null	0.286	Null	689.1	Null
2008	178	802	27.72	Null	0.367	Null	878	Null

Table 3: Meteorological data table after range check

Year	Day	Minute	Watertem p	Watertem p flag	Solarrad	solarrad flag	quantumra d	quantumra d flag
2008	178	800	27.68	3	0.257	3	647.6	3
2008	178	801	27.68	3	0.286	3	689.1	3
2008	178	802	27.72	3	0.367	3	878	3

File Edit View History Bookmarks Tools Help

http://www.mobilebaynep.com/mondata/rangestatus.cfm?stationid=703

Google Publish 2.71 Error Occurred While ...

Untitled Document Range Status

### Meaher Park Station Data Range Check Year 2008

This page displays only records with data out of range. Data in red is the specific reading out of range.  
 Click [here](#) to view range criteria.

**Hydrographical Data** [View Data](#)  
 All hydro data is within range.

**Meteorological Data**

Jday	Timedata	Precipitation	Air Temp	Solar Radiation	Quantum Radiation	Wind Direction	Wind Speed	Baro Pressure
165	1235	normal	normal	normal	2718.0	normal	normal	normal
165	1234	normal	normal	normal	2747.0	normal	normal	normal
165	1224	normal	normal	normal	2714.0	normal	normal	normal
165	1223	normal	normal	normal	2713.0	normal	normal	normal
165	1222	normal	normal	normal	2713.0	normal	normal	normal
165	1220	normal	normal	normal	2761.0	normal	normal	normal
165	1219	normal	normal	normal	2725.0	normal	normal	normal
165	1149	normal	normal	normal	2717.0	normal	normal	normal
165	1148	normal	normal	normal	2730.0	normal	normal	normal
165	1146	normal	normal	normal	2719.0	normal	normal	normal
165	1145	normal	normal	normal	2742.0	normal	normal	normal
165	1144	normal	normal	normal	2791.0	normal	normal	normal
165	1143	normal	normal	normal	2751.0	normal	normal	normal
163	1036	normal	normal	normal	2704.0	normal	normal	normal
163	1035	normal	normal	normal	2708.0	normal	normal	normal
163	1024	normal	normal	normal	2739.0	normal	normal	normal
158	1245	normal	normal	normal	2590.0	normal	normal	normal
158	1211	normal	normal	normal	2540.0	normal	normal	normal

Figure 1: Web page displays records with questionable data. The parameters flagged “2” are displayed in red.



#### 4. Implementation

Once the data are flagged, another program checks all the flags, four times every day. This program resides on our web server, and is written in Cold Fusion, a script language that can create interactive web applications. This program looks for any new flags marked “2”, or questionable/ suspect data. When it finds a flag marked “2”, it emails the technical support manager, the technicians, and the data manager. In the email, a link to the web page displays the exact record with questionable/suspect data. The actual reading of parameter flagged “2” is displayed in red (Figure 1).

The technicians, reviewing the questionable/suspect data, can make the final decision on the quality of the data after examining the following:

1. Physical Processes. Example: strong winds may result in unusually low tides
2. Possible malfunction of instruments. Example: accumulation of fouling organisms on sensor will affect the reading of dissolved oxygen
3. Biological Processes. Example: algae bloom may result in high dissolved oxygen reading.

**Dauphin Island Sea Lab**  
Environmental Monitoring Data  
Meteorological Data Update Screen  
Station Meaher Park  
Year 2008

**Update Individual Record(s)**  
Check the "Update" check box for those records that you want to update.

Year	Jday	Time	Quantum Radiation	Flag	Comments	Update	
2008	1	71	1238	2568.0	2		<input type="checkbox"/>
2008	2	88	1035	2554.0	2		<input type="checkbox"/>
2008	3	88	1035	2554.0	2		<input type="checkbox"/>
2008	4	88	1036	2581.0	2		<input type="checkbox"/>
2008	5	93	1157	2654.0	2		<input type="checkbox"/>
.....							
2008	91	165	1235	2718.0	2		<input type="checkbox"/>

**Update All Flags and Comments with the Same Values for the Above Records**  
(This function does not update any parameter values, such as Air Temperature, or Wind Speed. It only updates Flag or Comments. Leaving Flag or Comments blank will write null values into the database.)

Flag:

Comments:

Figure 2: Edit screen to revise parameter data value and flag value, or enter comments.

The technician can update the flag, or add comments to the record through a password protected website: <http://www.mobilebaynep.com/mondata/login.cfm>. Once logged on to the edit screen, the technician can revise a data-parameter value and its flag value, or enter a comment (Figure 2).

The metadata for the dataset has been registered with National Coastal Data Development Center (NCDDC), under the title “Mobile Bay Real-Time Continuous Environmental Monitoring” ([http://portal.ncddc.noaa.gov/approved\\_recs/org/disl/it/SeaLab/Monitoring/MonitoringData/MBEnvMonitoring.html](http://portal.ncddc.noaa.gov/approved_recs/org/disl/it/SeaLab/Monitoring/MonitoringData/MBEnvMonitoring.html)). In the “Data Quality Information” section of the metadata record, the review process is captured. Logging of the quality control process is recorded in Excel spreadsheet format by technicians. The data flags are available together with the data when web visitors download data from the DISL website (<http://www.mobilebaynep.com/mondata/disclaimer.cfm>) (Appendix 2). Figure 3 shows the relationship of different components in the range checking system.

Figure 3: Diagram of the Range Check System.

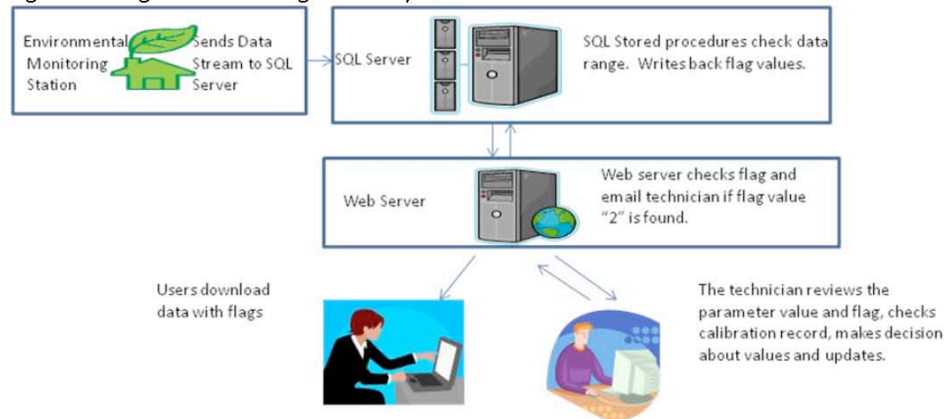


Figure 3: Diagram of the Range Check System.

## 5. LUMCON's Implementation

LUMCON's implementation is similar to DISL's with the exception of that the automatic range check is triggered whenever data are inserted into the database.

## 6. Discussion

Since we do not have the man-power for checking data manually due to the increased data volumes (one record every minute for meteorological data, and one record every 30 minutes for hydrographical data), automatic range checking has become the primary method for knowing when the data are out of range. When an out-of-range email is sent, the email will link to the exact record out of range. This information has helped the technicians to decide if instruments need calibration right away, or if the data is actually valid in some exceptional conditions. As the result of range checking, data are flagged. The data flags downloaded with the data from the DISL website offer users preliminary control over the quality of data. The next step in quality control will be implementing step-three procedures. We are currently working on step three, the implementation of an automated trend checking procedure as well as procedures to check data against nearby sites. Currently, we are servicing near real time data from all our stations to National Data Buoy Center (NDBC) every 30 minutes

([http://www.ndbc.noaa.gov/maps/Alabama\\_inset.shtml](http://www.ndbc.noaa.gov/maps/Alabama_inset.shtml)). NDBC performs checks of the data against nearby sites and informs the respective labs of the results. Quality Control (QC) of environmental monitoring data is an on-going process. Data without quality control is data with no significant potential usage. While implementing Quality Control such as range checking, we need to be aware that as environmental conditions change, and as situational needs change, the checking standard might change too. Only when we keep up with these changes and apply them in the standards, can we provide our users with updated good quality data.

### **Acknowledgements**

We want to thank Mike Dardeau, Holly Hebert, and technicians at the respective sites whose assistance made this report possible.

### **Reference**

- NOAA, 2004, First Workshop Report on the Quality Assurance of Real-Time Ocean Data, June 2004, 38pp, Stennis Space Center, MS.
- Small, D. Tamara, 2004a. Water Quality Data Management Procedures. In: CDMO NERR SWMP Data Management Manual. University of South Carolina, Georgetown, SC, pp. 53-163.
- Small, D. Tamara, 2004b. Meteorological Data Management Procedures. In: CDMO NERR SWMP Data Management Manual. University of South Carolina, Georgetown, SC, pp. 165-276.

Appendix 1: SQL stored procedure that checks the range of meteorological parameters

```
Begin Transaction

exec ('
update a set a.precip1Flag = case when b.precip1*2.54 >3 then 2
                        when b.precip1 is null then -9
                        when b.precip1*2.54 <=3 then 3
                        end,
a.airtemp1Flag= case when b.airtemp1 > 50 or b.airtemp1 <-10 then 2
                        when b.airtemp1 is null then -9
                        when b.airtemp1 >=-10 and b.airtemp1<= 50 then 3
                        end,
a.solarrad1Flag= case when b.solarrad1 >1.5 or b.solarrad1 < 0 then 2
                        when b.solarrad1 is null then -9
                        when b.solarrad1 >=0 and b.solarrad1 <=1.5 then 3
                        end,
a.quantumrad1Flag= case when b.quantumrad1 >2700 or b.quantumrad1 < 0 then 2
                        when b.quantumrad1 is null then -9
                        when b.quantumrad1 >=0 and b.quantumrad1 <=2700 then 3
                        end,

a.winddir1Flag = case when b.winddir1 > 360 or b.winddir1 < 0 then 2
                        when b.winddir1 is null then -9
                        when b.winddir1 >=0 and b.winddir1 <=360 then 3
                        end,
/*convert windspeed from knots to meter per second*/
a.windspeed1Flag= case when b.windspeed1/1.9438445 >30 then 2
                        when b.windspeed1 is null then -9
                        when b.windspeed1/1.9438445 <= 30 then 3
                        end,
/*convert inches of Mercury to millibar*/
a.bar_pressure1Flag= case when b.bar_pressure1/0.02953007 >1060 or
                        b.bar_pressure1/0.02953007 < 900 then 2
                        when b.bar_pressure1 is null then -9
                        when b.bar_pressure1/0.02953007 >=900 and
                        b.bar_pressure1/0.02953007 <=1060 then 3
                        end

from ['+'@station+'_'+'@Year+'_met_min] as a
inner join (
SELECT *
FROM    ['+'@station+'_'+'@Year+'_met_min] with (nolock)
WHERE   airtemp1Flag is null
) as b
on a.jday=b.jday and a.timedata=b.timedata')

select @errnum=@@Error, @RowCount = @@RowCount
if @errnum <> 0 GOTO sqlerror

Commit Transaction
```

Appendix 2: Data downloaded in .csv format from Dauphin Island Sea Lab website with data flags and a description of the aggregate quality flag:

1	Created on 06/26/2008										
2	Metereological Data Year 2008 5/1/2008 to 5/31/2008 Station Meaher Park										
3											
4	Metadata										
5	Year:Year										
6	Julian Day: Julian Day										
7	Time: 24 hour clock										
8	Air Temperature: Temperature measured in celsius										
9	Wind Direction: Wind direction in degrees from North										
10	Wind Speed: Wind speed measured in knots										
11	Barometric Pressure: Barometric pressure measured in inches of mercury										
12	Solar Radiation: Solar radiation measured in kW/m2										
13	Quantum Radiation (PAR): Quantum radiation measured in $\mu E/m^2/s$										
14	Precipitation: Precipitation total measured in inches										
15											
16	Aggregate quality flag										
17	0 or blank = quality not evaluated										
18	1 = bad										
19	2 = questionable/suspect										
20	3 = good										
21	-9 = null value										
22											
23	Year	Julian Day	Time	Air Temperature	Air Temp Flag	Wind Direction	Wind Dir Flag	Wind Speed (knots)	Wind Speed Flag	Barometric Pressure	Bar
24											
25	2008	122	1	17.71	3	132.6	3	6.004	3	30.04	
26	2008	122	2	17.72	3	152.9	3	6.283	3	30.03	
27	2008	122	3	17.74	3	133.1	3	5.309	3	30.03	
28	2008	122	4	17.73	3	136.2	3	6.001	3	30.03	
29	2008	122	5	17.7	3	132.2	3	5.522	3	30.03	
30	2008	122	6	17.72	3	147.1	3	5.807	3	30.03	
31	2008	122	7	17.71	3	141.5	3	6.137	3	30.03	
32	2008	122	8	17.7	3	139.9	3	5.96	3	30.03	
33	2008	122	9	17.7	3	138.9	3	5.353	3	30.03	
34	2008	122	10	17.7	3	138.2	3	6.309	3	30.03	
35	2008	122	11	17.69	3	135.7	3	6.764	3	30.03	

## CONVERTING DATA TO INFORMATION: COUPLING LAB-LEVEL DATABASE FUNCTIONALITY WITH PRIMARY LTER DATA ARCHIVING SYSTEMS

Adam M. Kennedy<sup>1</sup>, Suzanne M. Remillard<sup>1</sup>, Donald L. Henshaw<sup>2</sup>,  
Lawrence A. Duncan<sup>3</sup>, Barbara J. Bond<sup>1</sup>

<sup>1</sup> Department of Forest Science, Oregon State University, Corvallis, OR 97331, USA.

<sup>2</sup> U.S. Forest Service, Pacific Northwest Research Station, Corvallis, OR 97331, USA

<sup>3</sup> Orion Imaging, <http://orionimaging.net>, Portland, OR, 97202, USA

### Abstract

Developing and operating a data management program to support dynamic terrestrial and aquatic sensor networks is challenging. The database architecture needs to be robust and extensible, and must maintain flexibility in response to frequent changes in sensor array configurations in the field. The objective of this paper is to describe a database application developed for the Forest Ecophysiology and Ecohydrology Laboratory (FEEL) research program at the Andrews Experimental Forest in Oregon, USA. We discuss the fundamentals of a lab-level, web-based, and open source database application, and summarize the database architecture, methods of user-entered metadata, generation and storage of data mappings that provide the flexibility to handle changes in the incoming raw data streams, and methods to couple the lab-level database tables to the archival-level tables for seamless data flow and scheduled updating. This web-based database application enables small labs to handle large and streaming sensor arrays locally. The architecture is flexible and can adjust on-the-fly to changes in data file and field configurations. We detail a robust, user-friendly, and open source database environment that permits metadata generation and handling, low-level sensor tracking, dynamic data streams, general data processing, basic visualization, user-defined queries, and data routing to the primary long-term data repository.

**Keywords:** climate, sensor array, Andrews Experimental Forest, LTER, Forest Ecophysiology and Ecohydrology Lab, open architecture, environmental data

### 1. Introduction

Developing and operating near real-time terrestrial and aquatic sensor network data streams presents special challenges. The database architecture needs to be robust and extensible, and must maintain flexibility in response to frequent changes in sensor array configurations in the field. In competitively funded projects that are part of a larger umbrella network, such as the Long-Term Ecological Research (LTER) Network, the increased data stream poses significant challenges in that they are beyond the scope and intensity of what an information manager at an LTER site can accommodate. Most research labs lack the infrastructure and personnel to develop and or maintain a robust data management system. However, advances in data collection technology (Selker et al. 2006) and the associated increases in the volume of raw data streaming from the field, require new tools to handle these new data streams.

One solution is designing robust “lab-level” databases that are extensions of the primary Network database infrastructure. This type of hierarchical database schema provides users at the lab-level the flexibility to change sensor suites, add new sensors, and track more specific information relating to a sensor array apart from the primary database system, but still enables

easier subsequent integration to the Network-wide databases than if there were no prior coordination of schemas. Tools for these robust systems could include a secure multi-level and hierarchical user login system, functionality for user-entered metadata prior to sensor deployment, near real-time quality control of the data, basic visualization, and the ability for users to query the database for specific information without the need for expensive and memory intensive software installed on local or field computers. Such a lab-level system becomes critical to research success and longevity, and the ability to convert these raw data into useable information. Additionally, these systems can be designed to be tightly coupled with primary, long-term archiving systems, such as those represented within the distributed LTER sites nationally.

This paper focuses on the design and implementation of a lab-level system, which we defined here as the local lab domain operated at the lab-level and designed to have direct communication with the LTER site database domain. While it may generally operate at a small spatial scale, its contribution to the larger database domain remains integral to the success of the entire LTER site. It is at this level that the research programs funded by external grants operate and should maintain tight data relationships with the LTER site data managers.

This lab-level system consists of an online terrestrial database application that is currently deployed by the Forest Ecophysiology and Ecohydrology Lab (FEEL) at Oregon State University (<http://oregonstate.edu/feel>) and is being tested to handle near real-time data streams emerging from a small, steep-walled, forested watershed within the HJ Andrews Experimental Forest. The objectives of this paper are to summarize some of the techniques employed by this lab-level database ideology, and to serve as a building block for similar lab-level systems, as future sensor networks come online and create the need for solutions to growing data collections such as data storage, processing, and retrieval tools.

## 2. Methods

The climate and carbon research program within the FEEL increased in both data quantity and data type diversity during its four years of funding with the number of total continuous records approaching 50 million. This research program has nine plots, measures a full suite of environmental variables, and collects samples at intervals ranging from 1-minute to hourly. This example presents a clear need for robust applications to convert raw data points into useful information. The relational database selected for the FEEL application was MySQL (<http://mysql.com>), which provides a free, robust, relatively easy to use, and open source computing architecture. The platform permits advanced relational database functionality and can be coupled to other commercial database infrastructures using readily available MySQL ODBC drivers. The metadata entry component of this package was built using the PHP programming language with user authentication routed through online hypertext transfer protocol over a secure socket layer connection (HTTPS), which is used to indicate a secure HTTP connection. The PHP layer allows a user to avoid interacting with the database directly, and permits metadata entry and data processing to occur over a HTTPS connection through dynamic drop down lists and text boxes that insert the required data into the database in the proper format. Each piece of this application was developed to perform a specific task, but integrating the pieces has resulted in a robust application capable of managing a sensor network. While the database schema is not included in this paper, a summary of the FEEL application online metadata and data import procedures is shown in Figure 1.

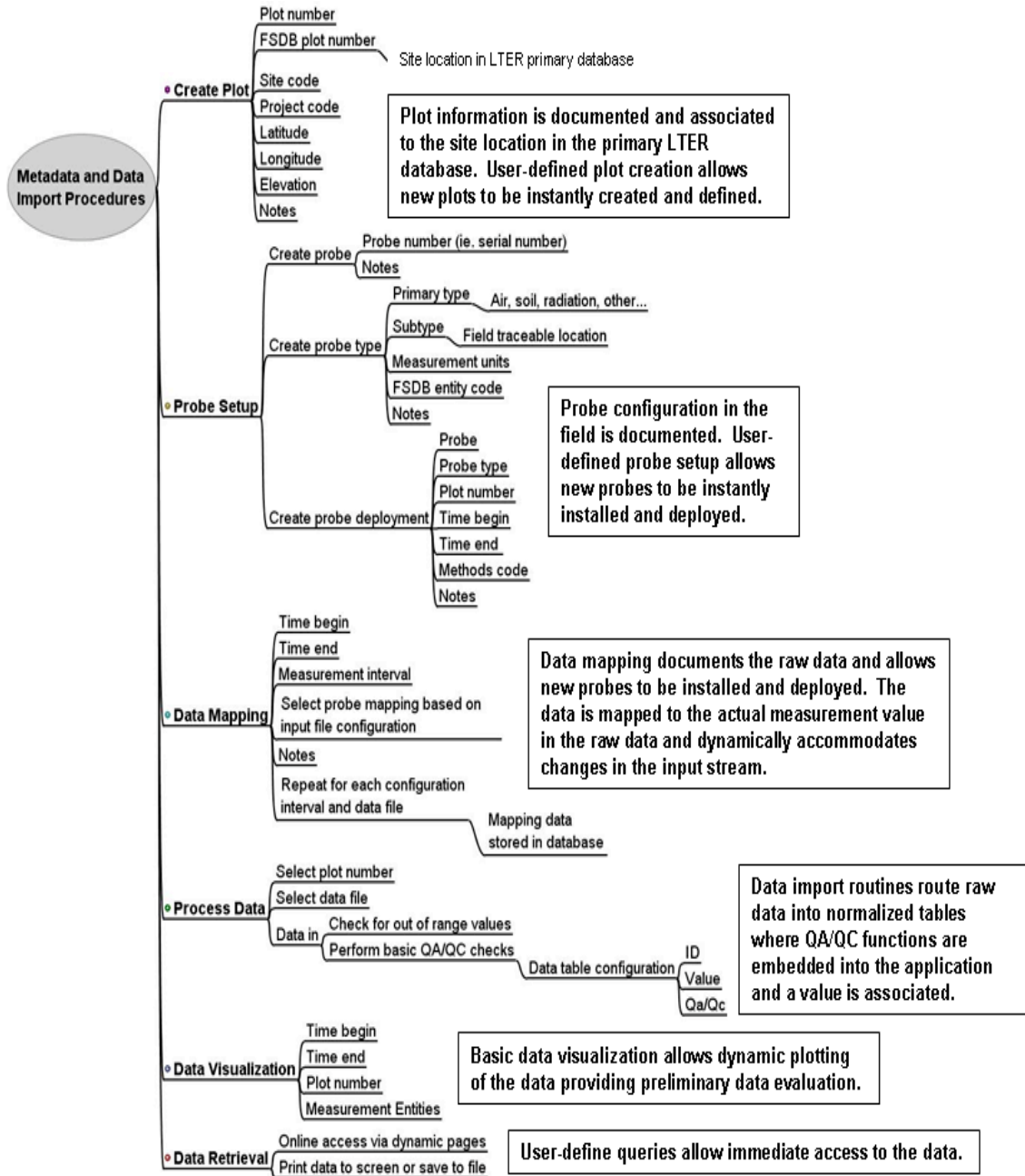


Figure 1. The FEEL database application was built to help manage detailed probe movement and calibration. The metadata and data import procedures and functionality include a secure user login system, user-entered metadata capabilities, near real-time quality control of the data, basic visualization, and the ability for users to query the database for specific information. The metadata entry component of this package allows the user to create plots, describe probe and probe deployment setup, and to handle dynamic field sensor configurations with on-the-fly data mapping setup tools. The metadata and data tables are linked such that the dynamic import of data streams accommodates changes in the input stream. Additionally, data visualization and retrieval functions allow immediate access to the data enabling preliminary data evaluation.

Defining valid and correct data mappings is an essential step of data processing, with the quality of the data mappings being critical to the quality of the extracted data. Our development



team has engineered a command line tool, *df\_info* (a mnemonic for "data file information"), to assist in the initial assessment and analyses of raw data files, in preparation for defining data mappings. *Df\_info* utilizes the PHP command line interface (CLI) (<http://php-cli.com>) framework to allow data files to be rapidly parsed, extracting all observation intervals that are present, while demarcating the line numbers on which changes in instrumentation occur. Once data mapping occurs, the data file is then ready to be processed with the FEEL application. Data mappings need only to be revised after a known configuration change has occurred.

The novel idea of data mapping allows for dynamic import of data streams into the application even when changes in the input stream exist mid-file. We tested the accuracy of this system by computing the fraction of import success over import failure. Import success rates are logged and available for retrospective analysis. On-the-fly plotting functionality is in place for preliminary data evaluation. The plotting module employs an off-the-shelf plotting application (<http://jgraph.com>) coupled to the FEEL application and, as of this publication, is used only for preliminary evaluation or broad data quality checks. To connect the FEEL lab-level database to the primary LTER site SQL Server database, we employed Microsoft's SQL Server Integration Services (SSIS), which is embedded in SQL Server 2005 Management Suite (Figure 2). This tight coupling to the LTER site database at the lab-level promotes semi-automatic data warehousing and reduces the time required to continually update continuous sensor array data from the field.

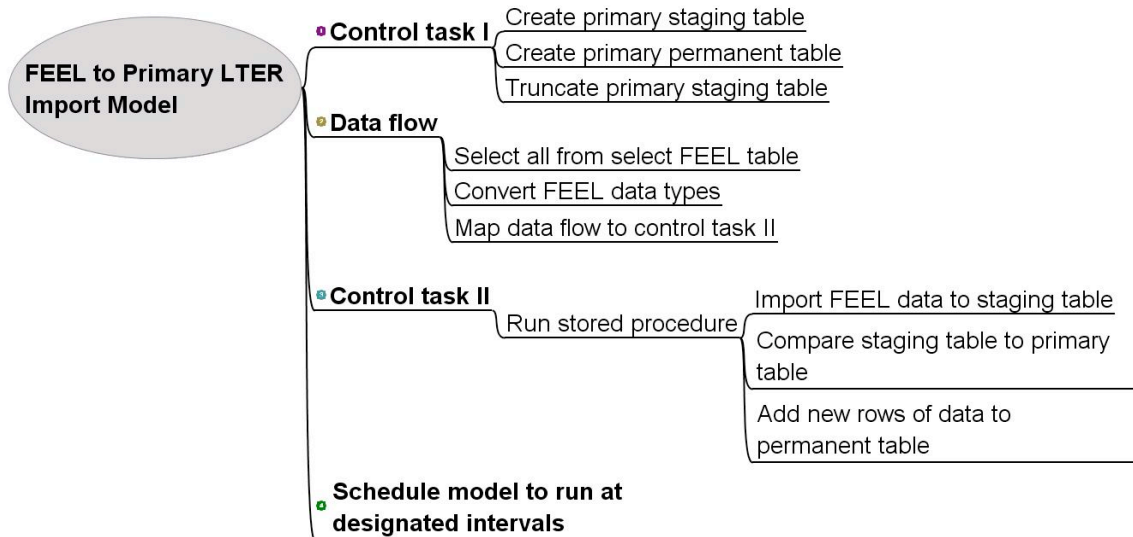


Figure 2. The FEEL database uses an SQL Server Integration Services algorithm to execute communication to the primary LTER database for long-term data archiving.

### 3. Results

Implementing the success of this system required several months of intensive but intermittent focus. The resulting database application achieves our objectives by handling the FEEL data needs; metadata (including sensor calibration notes, replacement, type, serial number, units, dates deployed, etc.) are accessible, individual sensors may be tracked in the field, input

file configurations can be appropriately mapped so that the database import features will know how to handle each record, and basic QA/QC functions are embedded into the application.

Attention was given to processing speed, long-term storage, writing optimized queries to insert and route data to their appropriate tables, and to return data to users via online data access portals. The user-entered metadata are stored in tables within the FEEL database and metadata are available through user-defined queries via a web portal. Data import functions route the raw data into normalized tables and a QA/QC field is associated with each record.

A general QA/QC model was developed for timestamp and min/max range checking as a function of month within the database environment that leaves data as is, but flags each questionable or missing value with an assigned value. Evaluating the data as a function of time and value instead of an absolute value tightens the band into which good values must exist during the QA/QC procedure, raw data are left joined to a “calendar” table so that all timestamps are included and missing raw records are inserted as null values.

Data mappings assigned to each raw input file result in an import success rate approaching 99.9%. When import failures occur, they are generally due to a temporal change in sensor configuration, or a duplicate timestamp due to data logger failures. While these count as “import failures”, the former is isolated to the record where the configuration change occurred, and the latter is considered advantageous, in that our automated import routine will not import duplicate timestamps, thus preserving the integrity of our database. Files with no changes in sensor configurations typically have a 100% import success rate. More detailed error analysis will be performed in later releases of the FEEL database.

The FEEL application is flexible in that little additional programming is required to add new database fields as metadata needs evolve and portable in that it will run on a “localhost” or remote server with proper configuration. For our localhost environment, we currently employ XAMPP, which is an easy to install Apache distribution containing MySQL, PHP, and Perl (<http://www.apachefriends.org>). A complete installation using documentation averages less than one day. If the application were to be used by another lab, the main requirement is that the input files need to be comma delimited, and the first five fields must include (in this order) – array number (which describes the sampling interval), site id, year, day of year, hour/minute – with the remaining fields being the observed variables.

#### **4. Discussion**

Lab-level tools need to maintain the strict metadata protocol built into the primary LTER database, but also embrace the abundance of raw data streaming from the field and have the ability to convert the raw data streams into a complete set of information used to answer questions from all levels of students, the research program personnel, and policy makers. While the LTER network and many LTER sites have developed integration tools and techniques for core datasets at the inter-site level, like ClimDB/HydroDB (<http://www.fsl.orst.edu/climhy>), and the site level (GCE (Sheldon 2003)), lab-level components, derived from the recent onset of data rich and externally funded research programs, are missing from the LTER tool suite.

New products, both commercial and open source, such as DataTurbine (<http://dataturbine.org>), Antelope (<http://brtt.com>), SensorBase (<http://sensorbase.org>), and GSN (<http://gsn.sourceforge.net/>) are being developed to provide means to handle large amounts of streaming data. Commercial products were not considered financially feasible for the FEEL lab and most of the open source applications were still in development when the lab-level application was being developed. The FEEL lab decided to develop a custom tool with the intention of creating a congruent system that could be easily adapted by other labs. An

immediate advantage of this application is its ability to integrate directly with the existing LTER site database. Improved interoperability between the major data management tools and existing data management programs could improve the effectiveness of these tools.

There are two fundamental ideas presented in this paper that make tools like the FEEL database necessary. First, management of sensor array data is critical for maintaining and assuring data quality in near real-time. Preserving low-level metadata regarding collected attributes, array configurations, probe placement and sensor calibration is critical to track field sensor array history in long-term studies. Raw data streams and low-level metadata must also be available in a useable format to researchers for regular quality checks and sensor calibration history. Secondly, lab-level databases at the LTER sites can accommodate intensive studies and manage associated sensor arrays that are outside the scope of the primary information management system, but still integrate final data into the system

As presented in this paper, it is necessary that the lab-level system be managed independently from the LTER primary database but must still function in conjunction with the long-term data archive to gain benefits from that system such as metadata-driven data validation, compliance with network-wide metadata standards, generation of Ecological Metadata Language, and participation in network-wide databases.

### **Acknowledgements**

We would like to acknowledge the Orion Imaging programming staff for their consultation and interest in this project and the system administrators at Oregon State University Central Web Services and the OSU College of Forestry. Gabriel Shea of Idea Pivot provided the initial expertise needed to complete the FEEL to LTER database coupling. Finally, we would like to thank H. Barnard, N. Czarnomski, J. Gabrielli, R. Hopson, Z. Kayler, C. Phillips, T. Pypker, S. SanRamoni, B. Wilson, and E. Wyckoff who have helped test the FEEL application during the implementation of the FEEL system. Funding for this work was provided by the NSF “A Wireless Network of Battery-Free Sensors for Atmosphere-Biosphere Studies in Complex Environments” grant (DBI0529223) and the NSF Andrews LTER grant (DEB0218088).

### **References**

- Selker, J.S., L. Thévenaz, H. Huwald, A. Mallet, W. Luxemburg, N. van de Giesen, M. Stejskal, J. Zeman, M. Westhoff, 2006. Distributed fiber-optic temperature sensing for hydrologic systems. *Water Resources Research* 42: W12202 1 – 8. doi: 10.1029/2006WR005326.
- Sheldon, W.M., 2003. Presentation: Software tools for automated metadata creation, metadata-mediated data processing and quality control analysis – real time processing solutions for real-time data. 2003 LTER All-Scientists Meeting. Sept. 18-21, Seattle, WA.

## DETECTING SENSOR FAILURES IN ECOLOGICAL SENSOR NETWORKS

O Langman<sup>1</sup>, PC Hanson<sup>1</sup>, SC Carpenter<sup>1</sup>, K Chiu<sup>2</sup>, YH Hu<sup>3</sup>

<sup>1</sup>University of Wisconsin – Madison, Center for Limnology, 680 North Park Street, Madison, Wisconsin 53706 USA

<sup>2</sup>SUNY Binghamton – Department of Computer Science, P.O. Box 6000, Binghamton, NY 13902-6000

<sup>3</sup>University of Wisconsin – Madison, Electrical and Computer Engineering, 1415 Engineering Drive, Madison, Wisconsin 53706 USA

### Abstract

We present and evaluate a Bayesian method ('Surprise Theory') for detecting changes indicative of sensor malfunction within data measured by autonomous sensor networks. Surprise Theory is evaluated under simulated conditions representative of known anomalies to test its detection capability. Real world performance is evaluated by comparing expert classification of anomalies to surprise model classification of anomalies within a dataset comprised of sensor data obtained from a wireless sensor network measuring thermal profiles, chemical variables, and meteorological conditions in northern temperate lakes. Within this two year dataset comprised of a diverse range of sensors and containing many distinct types of sensor malfunctions, 91.5% of the errors classified by experts were correctly classified by Surprise Theory using conservative parameters. We conclude that Surprise Theory has potential uses as a screening tool to help users identify plausible problems in streams of sensor data in ecology.

**Keywords:** event detection, ecological sensor networks, Surprise Theory, sensor malfunction

### 1. Introduction

Sensor networks are used increasingly in the ecological sciences due to their capacity to regularly and reliably gather data for a variety of applications. Early results from such installations (Collins et al. 2006, Hart and Martinez 2006, Effler et al. 2002) and the promise of long term, high resolution data suggest an expanding reliance on sensor networks to provide critical data for basic and applied research. The lessons learned from early sensor deployments have led to improved platform designs in terms of both ease of deployment and reduction in cost, ultimately decreasing activation energy for the deployment of new networks that show promise for new science (Porter et al. 2005).

Despite improvements in implementation and deployment, the long term maintenance needed to operate a sensor network still requires significant and sustained human involvement. Since benefits derive from constant data flow to capture transient important events (Daly et al. 2004, Porter et al. 2005), interruptions in communication with the remote platform and failures in sensors must be dealt with quickly. Thus costs of monitoring and maintaining networks are high. Furthermore, sensor malfunctions are often difficult to characterize without advanced knowledge of the variables and system being monitored.

Automating sensor monitoring requires a method that is capable of (1) characterizing real time observations, (2) recognizing deviations from past performance indicative of sensor malfunction, (3) distinguishing sensor malfunctions from normal perturbations, and (4) performing reliably with little maintenance. We propose that a Bayesian technique developed to

mimic human capacity for focusing on sensed events, Surprise Theory (Itti & Baldi 2006, Itti & Baldi 2005), can be applied with modification to achieve these goals and provide a reasonable level of automation in the monitoring of ecological sensor networks. This algorithm characterizes a sensor trace by incorporating measurements into a Bayesian prior, and identifies anomalies by comparing a new datum against the prior. An unexpected, i.e. surprising, datum will require the prior to be significantly modified, indicating an event within the sensor trace.

Here we apply Surprise Theory to detect anomalies in data obtained from lake sensor networks. We simulate conditions to test its detection capability under a variety of known variance, trends, and other conditions in which anomalies have been embedded. We use Surprise Theory to detect anomalies in real sensor data and test it against classification by experts. Surprise Theory showed an ability to detect a variety of anomalies under a variety of conditions, suggesting that it will be a useful tool for the maintenance of ecological sensor networks.

## 2. Methods

Surprise Theory is designed to focus on immediate changes within sensor traces and is built upon Bayesian principles. Past observations will be encompassed within a prior distribution which represents our knowledge of the sensor trace. A new observation will be added to this set of knowledge, creating a modified distribution, the posterior. If the new observation was expected, the two distributions will remain similar, but if the observation was unexpected, the

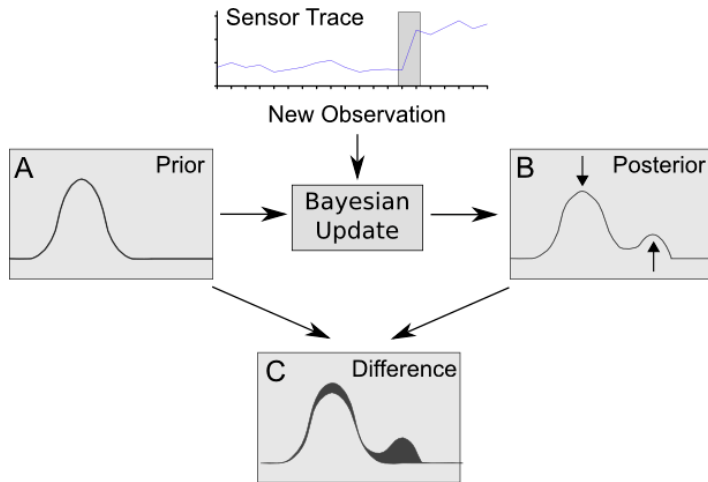


Figure 1: A simplified visual representation of surprise theory. The posterior shown (B) is generated via a Bayesian update after an observation was recorded that was not expected within the context of the prior (A), which was generated as a function of previous observations. The difference between the two distributions (C) is interpreted as a measure of how unexpected the observation was with respect to the previous observations and termed 'surprise'.

distribution shapes will be quite different (Figure 1). The measurement of this difference is termed 'surprise'.

*Model parameters* - There are two parameters within a surprise model that must be determined before implementation. The  $\alpha$ -parameter ranges from  $0 < \alpha < 1$ , and it can be thought of as a term that describes the memory of the system. Increasing  $\alpha$  results in a surprise model that remembers past states of the trace for a longer period of time. The second parameter is the choice of probability density function (PDF) used for the prior and posterior distributions. While not in actuality a Poisson process, sensor traces from environmentally sensed data often display a high degree of stochasticity such that a gamma density function performs well as the

PDF in the majority of cases. However, any PDF that has a calculable Kullback-Liebler divergence and Bayesian update is a potential candidate for use (Itti & Baldi 2006).

*Performance Analysis* - Model performance was analyzed by examining a combination of constructed signals and evaluating performance against real-world data. In both sets of trials, surprise models used a gamma PDF and an  $\alpha$ -value of 0.3. All models were initialized with

arbitrary values for the shape and scale parameters ( $\theta = 1.0$ ,  $\kappa=1.0$ ). The constructed sensor trace in Figure 2 is composed of a signal generated from a normal distribution with ambient noise ( $N(\mu=10, \sigma=0.5)$ ), punctuated by deviations from that sampled distribution. The deviations represent basic variations that correspond to sensor malfunctions as observed in sensor traces, specifically shifts in mean and both increases and decreases in variance.

Surprise models were tested against two years of sensor data to test overall detection capability and performance in a real-world sensor network installation. The sensor network was comprised of 6 sensor platforms with a suite of sensors obtaining data ranging from thermal profiles to dissolved oxygen in northern temperate lakes. A total of 69 sensor malfunctions were recorded in the metadata for the system, with causes ranging from total sensor failure to sensor movement within the environment. Surprise models used a gamma PDF and an  $\alpha$ -value of 0.3. All models were initialized with arbitrary values for the shape and scale parameters ( $\theta = 1.0$ ,  $\kappa=1.0$ ). Each sensor trace was analyzed independently and aggregated to produce a confusion matrix (Figure 5A). Performance within this dataset was analyzed for different ranges of  $\alpha$  (Figure 5B). Two specific traces and model response (Figures 5 & 6) are presented as examples of particular responses to sensor malfunctions.

### 3. Results

Many of the sensor malfunctions recorded in the real-world observations exhibit localized shifts in mean and variance. In Figure 2, a surprise model is tested against these common variations. While small shifts in the mean (Figure 2, A) are readily detected, the shift in general must exceed the variance of the system to generate surprise. This holds

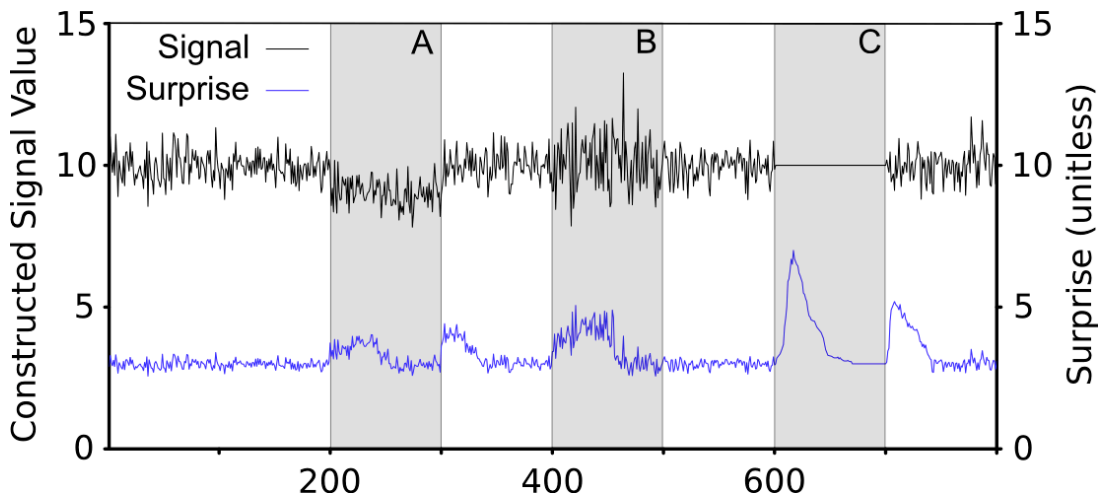


Figure 2: Surprise performance in the presence of noisy data. Surprise model is using a gamma distribution with  $\alpha = 0.3$ . Constructed signal data sampled from  $N(\mu = 10, \sigma = 0.5)$ . Highlighted areas exhibit modifications to the generated data. A) Decrease in the mean to  $\mu = 9$ , B) Increase in the variance to  $\sigma = 1$ , C) Decrease of the variance to  $\sigma = 0$ . Note that surprise is generated both during the alterations to the original signal as well as the return to the signal due to the small  $\alpha$  value.

true in the presence of more complex traces that exhibit regular fluctuations. Note that the return to normality within the trace also causes surprise to be generated and if the altered state persists over a long duration it may appear as two distinct surprise events. Changes in the variance are also readily detected (Figure 2, B and C). While a simple threshold analysis detects the increase in variance in B in this example, a change in the variance of a trace exhibiting with noise would

have not been detectable with thresholds, or in the presence of a reduction in variance because the signal would not necessarily have exceeded the defined bounds used in range checking.

The real-world data used to evaluate Surprise Theory represents the potential range of variability within sensor traces both in that it exhibits large amounts of pattern within the data that is not indicative of malfunction and contains a wide range of sensor malfunctions (Figures 3 and 4).

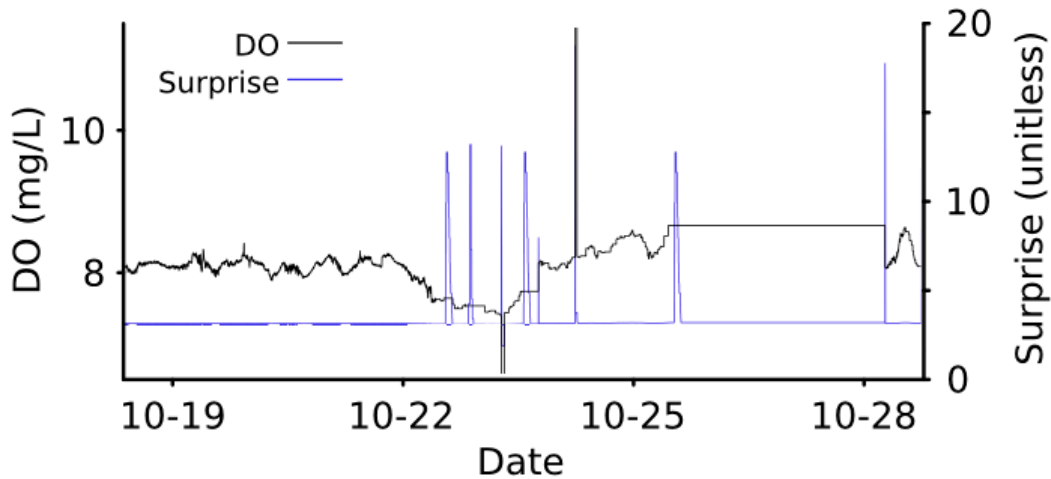


Figure 3: Dissolved oxygen trace from a sensor located in Lake Mendota, Wisconsin, USA. When the sensor failed to make measurements, it returns the value of the previous measurement. There are several instances before the sensor failed entirely where it failed for short periods of time. Due to the very high sampling frequency (1 sample min<sup>-1</sup>), the malfunction was not noticed immediately.

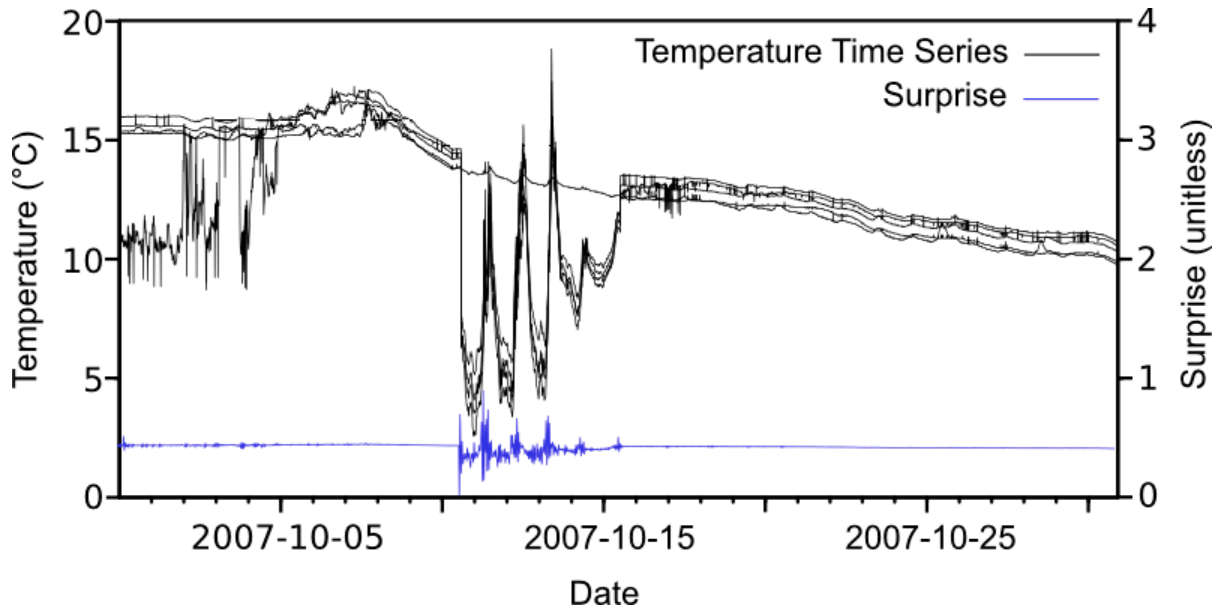


Figure 4: Temperature time series data from October 2007. A sensor malfunction occurs on the 10th of October. Surprise is generated even though the model had been trained on data that contained thermal stratification. The surprise model incorporated a total of 20 depths, 5 of which are shown.

A tradeoff exists regarding the choice of  $\alpha$ -parameter. Higher values result in a larger fraction of the sensor malfunctions being detected, but false positives increase exponentially (Figure 5B). A conservative value of  $\alpha$  ( $\alpha=0.3$ ) is the solution for a function that minimizes false positives and maximizes detection, putting equal weight on both goals. The resulting performance is a detection rate of 91.5%, with 62 false positives.

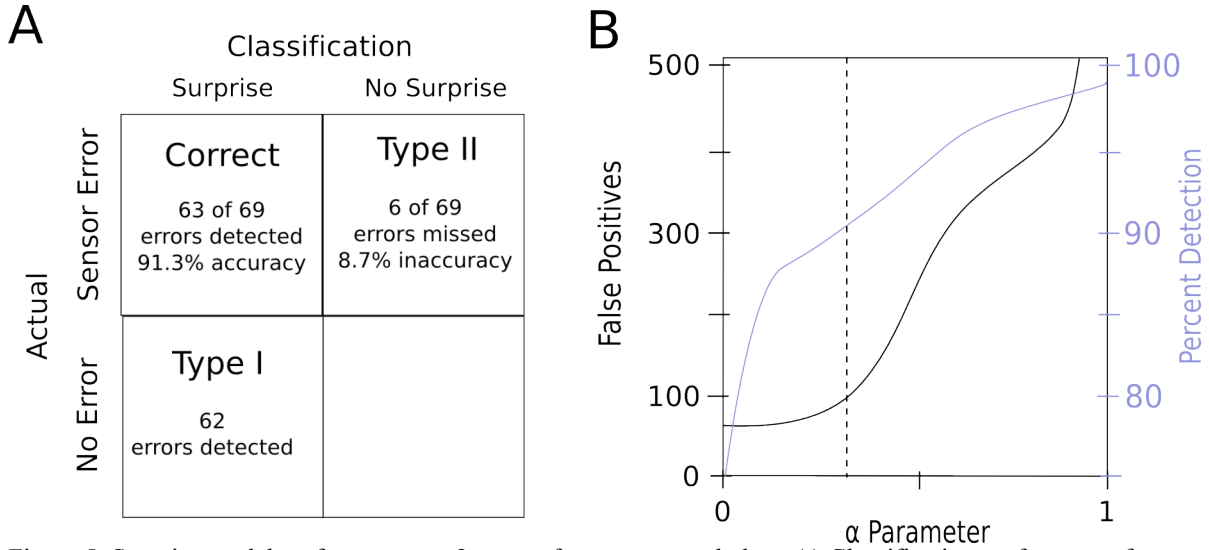


Figure 5: Surprise model performance on 2 years of sensor network data. A) Classification performance for one specific value of alpha (0.3). B) Classification performance over the range of potential alpha-values. The dashed line indicates the alpha value at which A was calculated.

Focusing on immediate patterns within sensor traces eliminates the ability to detect gradual shifts within sensor readings. Within the framework of surprise theory, a gradual change in the mean or variance of a system will not greatly alter the posterior distribution with each time step, producing little change between the posterior and the prior (Figure 4). Sensor malfunctions that exhibit this gradual shift pattern (primarily sensor drift) will only be detectable by altering the scale at which the model is sensitive. Subsampling the original sensor trace can be applied to overcome this limitation.

#### 4. Discussion

The application of Surprise Theory is motivated by our need to quantify events within data obtained from diverse installations of networked sensors. The sensor traces produced by ecological sensor networks have several distinct traits that are persistent across most installations despite their often differing motivations: (1) measurements occur in discrete time, (2) observations display a high degree of stochasticity, and (3) there often exists a high degree of periodicity within the data across several different temporal scales (Green et al. 2005, Porter et al. 2005). Discrete time measurements (1) are exploited in that they guarantee that units of surprise are comparable across time steps. Stochasticity and periodicity (2)(3) are incorporated into the PDF. These methods of incorporating consistent properties into the surprise model lead to: minimal training and knowledge of the system are needed to implement a surprise model, the models exhibit invariability with respect to scale, and the models perform well in the presence of diverse real-world data. Despite the variability in conditions, the algorithm was configured identically for all sensors in the test dataset, indicating a strong capacity for generalization. This



has obvious benefits in terms of ease of deployment and use, but ultimately generalization is important because the algorithm can readily be applied to systems with unknown properties.

Surprise Theory is intended as a way of focusing human attention on potential sensor malfunctions, not as a way of replacing human monitoring entirely. The misclassification errors within the real-world analysis (Figure 5) are complex and perhaps oversimplified in the confusion matrix. In addition to non-recorded actual malfunctions, sudden ecological events can potentially generate surprise. Within the false positives, 4 of the events correlate with a major storm event that triggered surprises for wind speed and direction sensors. The wide range of possibilities in misclassification are problematic in that many of them appear similar to malfunctions. Human oversight is still needed to distinguish between malfunctions and other surprising events. Nonetheless, Surprise Theory can be used to identify events worthy of human attention.

Detecting events within sensor networks is a goal that will see significant future research. While detecting sensor malfunctions is useful in the normal operation and maintenance of these networks, the capacity for detecting ecological events is the obvious next step. The surprise model as presented here detected several ecological events, but since most ecological events do not occur at the same rates relative to the speed of sensor measurements, the algorithm will not detect most of them as it is currently configured. Distinguishing between ecological events and sensor malfunctions still requires human guidance. Beyond simply detecting events, integrating the detection algorithms at the sensor platform or even sensor level could lead to even more autonomous systems. Integrating them at the platform level would let non-networked systems respond to ecological events by increasing sampling rates or activating sensors that are energetically costly to operate. Detection of sensor malfunctions could be followed by an automated recovery process. Sensor networks have already drastically increased our ability to achieve high frequency sustained sampling, but many of these future developments will allow an even greater degree of flexibility, while improving our capability to answer ecological questions.

### **Acknowledgments**

We would like to thank the Northern Temperate Lakes Long Term Ecological Research (NTL-LTER) program for providing funding for this research. Members of both LTER and GLEON provided useful input.

### **References**

- Collins, S.L., Bettencourt, L.M.A., Hagberg, A., Brown, R.F., Moore, D.I., and Bonito, G.D. 2006. New opportunities in ecological sensing using wireless sensor networks. *Frontiers in Ecology and the Environment* 4(8): 402–407.
- Daly, K.L., Byrne, R.H., Dickson, A.G., Gallager, S.M., Perry, M.J., and Tivey, M.K. 2004. Chemical and biological sensors for time-series research: Current status and new directions. *Marine Technology Society Journal* 38: 121-143.
- Effler, S.W., O'Donnell, D.M., and Owen, C.J. 2002. America's most polluted lake: using robotic buoys to monitor the rehabilitation of Onondaga Lake. *Journal of Urban Technology* 9: 21-44.
- Green, J.L., Hastings, A., Arzberger, P., Ayala, F.J., and Cottingham, K.L. 2005. Complexity in ecology and conservation: mathematical, statistical, and computational challenges. *BioScience* 55(6):501–510.

- Hart, J.K., and Martinez, K. 2006. Environmental Sensor Networks: A revolution in earth system science? *Earth-Science Reviews* 78: 177-191.
- Itti, L. and Baldi, P. 2006. Bayesian surprise attracts human attention, In: *Advances in Neural Information Processing Systems*, Vol. 19. Cambridge, MA:MIT Press.
- Itti, L. and Baldi, P. 2005. A principled approach to detecting surprising events in video. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Volume 1. CVPR.
- Porter, J.H., Arzberger, P., Braun, H-W., Bryant, P., Gage, S., Hansen, T., Hanson, P., Lin, F.P., Lin, C.C., Kratz, T., Michener, W., Shapiro, S., and Williams, T. 2005. *Wireless Sensor Networks for Ecology*. *Bioscience* 55: 561-572.
- Venables, W.N. and Ripley, B.D. 2002. *Modern applied statistics with S*. Springer, New York, 495pp.

## USING METADATA FOR LOADING AND QUERYING HETEROGENEOUS SCIENTIFIC DATA

Ben Leinfelder<sup>1</sup>, Jing Tao<sup>1</sup>, Duane Costa<sup>2</sup>, Matthew B. Jones<sup>1</sup>, Mark Servilla<sup>2</sup>, Margaret O'Brien<sup>3</sup>, Chad Burt<sup>3</sup>

<sup>1</sup> National Center for Ecological Analysis and Synthesis, University of California Santa Barbara

<sup>2</sup> Long Term Ecological Research Network, University of New Mexico

<sup>3</sup> Santa Barbara Coastal LTER, University of California Santa Barbara

### Abstract

The Ecological Metadata Language is an effective specification for describing data for long-term storage and interpretation. When used in conjunction with a metadata repository such as Metacat, and a metadata editing tool such as Morpho, the Ecological Metadata Language allows a large community of researchers to access and to share their data. Although the Ecological Metadata Language/Morpho/Metacat toolkit provides a rich and seamless data documentation mechanism, current methods for retrieving metadata-described data can be laborious and time consuming. Moreover, the structural and semantic heterogeneity of ecological data sets makes the development of custom solutions for querying them prohibitively costly. The Data Manager library leverages the Ecological Metadata Language to provide automated data processing features that allow efficient data access, querying, and manipulation without custom development. The library can be used for many data management tasks and was designed to be both extensible and easy to incorporate in existing applications. In this paper we describe the motivation for developing the Data Manager library, provide an overview of its implementation, illustrate ideas for potential use by describing several planned and existing deployments, and describe future work to extend the library.

**Keywords:** Heterogeneous data, metadata, data query, synthetic data

### 1. Data Heterogeneity (Introduction)

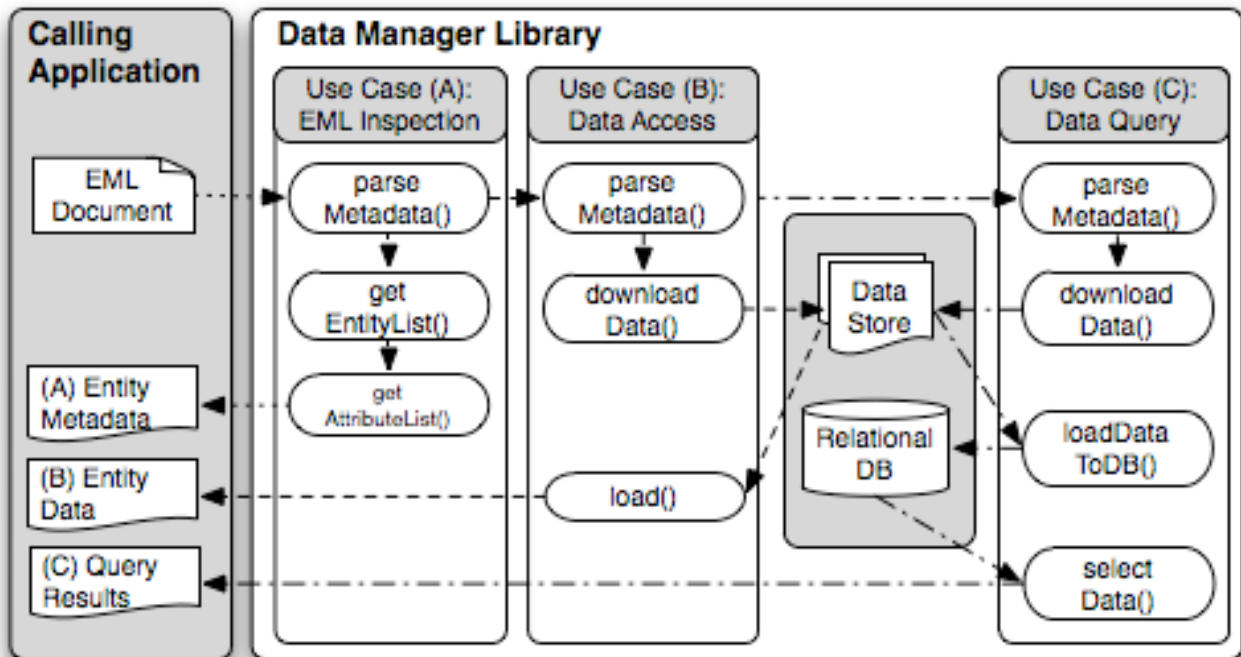
Even a cursory examination of the ecological data housed in the Knowledge Network for Biocomplexity (KNB) data repository and similar repositories such as the National Biological Information Infrastructure (NBII) Metadata Clearinghouse reveals the wide variety of data structures used by researchers to capture their observations (Jones 2006, Andelman 2004, Parr and Cummings 2005). The strength of the KNB model is that it easily supports data storage without prescribing a particular serialization mechanism or imposing structural constraints on data files. Ecological Metadata Language, or EML (Feagraus et al. 2005) allows data owners to preserve their original format by *describing* it rather than *conforming* to any given one.

Although the KNB is able to provide an effective storage solution for heterogeneous data, accessing that data has heretofore been an *ad hoc* process of downloading original data files and processing them manually. Managing these datasets can become increasingly taxing on resources – human and technological and creating custom individual solutions on the desktop encourages unscripted data management and irreproducible analysis processes.

## 2. Data Manager Library

The purpose of the EML Data Manager Library is to provide a common software library for parsing EML metadata and using the structural information about the data to create relational database structures, load associated data into the resultant database, and support query and selection operations on the data (Figure 1). Thus, without incurring the costs of custom database development, researchers gain relational database access to scientific data that may (and usually does) have an entirely different native format (i.e. a text file). The 'eml-dataset' and 'eml-dataTable' modules provide the explicit descriptors needed to allow both humans and software to accurately process the data structures.

The Data Manager Library includes an Application Programming Interface (API) so that applications can incorporate the library for their own use. Allowing multiple software applications to re-use the data manager library simplifies development and reduces the burden on application developers.



**Figure 1** Highlights the major operations supported by the Data Manager Library. Three use cases are illustrated. A host application that incorporates the library first makes a request to parse an EML document (A, B, C). The library optionally explores the metadata (A), downloads the data to the host data store (B, C), and creates backing tables in an associated relational database (C). After these tables are created, the library loads data into the database (C) and processes database queries on behalf of the application (C).

### 2.1 Parsing and Loading

The Data Manager library relies on well-documented data in order to support even the most basic query requests. The eml-dataset schema module supports such expressiveness, but it is incumbent upon the authors of said metadata to provide complete and accurate information. Metadata documents that do contain quality dataset descriptions are perfect candidates for use via the Data Manager.

EML documents can be accessed in a variety of ways by the Data Manager parser. Commonly they are served from a remote storage system such as Metacat that handles the metadata and the data storage as well as providing versioning capabilities. Datapackages represent a parsed version of EML's dataset element, a collection of data entities that in many cases refer to tabular data structures. These entities are further described by the attributes they contain. Each entity maps to a table in the backing relational database used by the Data Manager. The attributes of each entity correspond to the columns of those backing tables.

After the metadata have been successfully parsed and the backing table[s] created in the relational database, the actual data are retrieved and the table[s] populated. Loading the data also relies on details captured in the eml-dataTable module where specific field and record delimiters can be identified and where other complex structures can be described. The Data Manager is intended to support the variety of custom formats that can be fully described using EML.

The library supports many methods for accessing data. These include file transfer protocol (ftp), hypertext transfer protocol (http), Storage Resource Broker (SRB), and even the local file system. The Data Manager gathers data automatically with native support for the webservice of the EcoGrid system (<http://seek.ecoinformatics.org/Wiki.jsp?page=EcoGrid>).

## 2.2 Queries

Using the Data Manager library, a client application can assemble a query against the tables in a data package, execute that query, and retrieve the results. This allows client

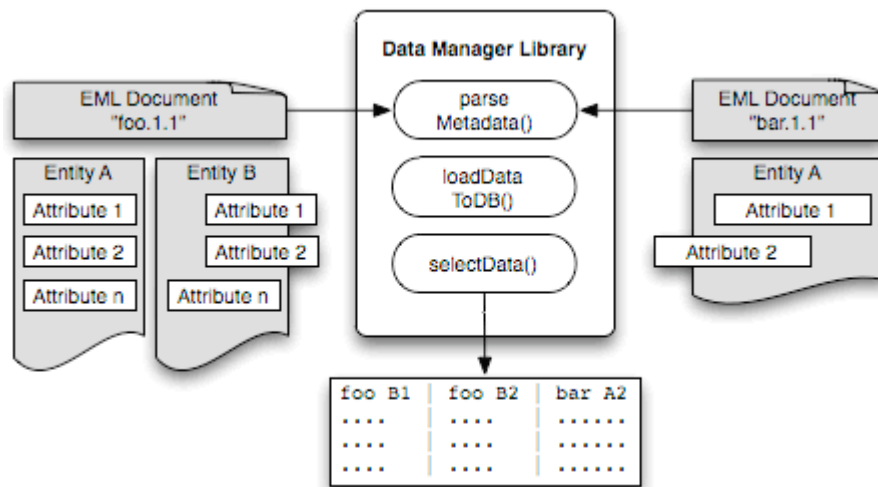


Figure 2 A query across multiple EML data packages (foo.1.1 and bar.1.1). The associated tables (Entity B and Entity A, respectively) are joined. Attribute 1 and Attribute 2 from foo's Entity B (foo B1, foo B2) and Attribute 2 from bar's Entity A (bar A2) are included in the query results. The Data Manager is metadata-driven and accommodates arbitrary table schemas such that entities can be structurally dissimilar.

applications and analysis systems like Kepler to only access the subset of a data, rather than the whole table. The Data Manager API provides methods for specifying compound selection criteria such that both the height and width of the selection (records and columns, respectively) can also be explicitly defined.

Related data may often be stored within the same Datapackage, but there is no prohibition on inter-Datapackage selection operations. Join and Union operations are supported

across multiple Datapackages to produce collaborative, truly synthetic results based on a variety of source metadata documents (Figure 2).

### **2.3 Table management**

Because one of the underlying goals of developing the Data Manager library was to alleviate the data management burden on the analyst, a built-in table management feature handles table creation, attribute naming, data lifespan tracking and storage space allocation. When accessed via the API, full and original metadata (names, descriptions, units, etc.) of the queried Attributes are available regardless of any name-mangling performed for RDBMS-specific database compliance. This provides a seamless system for simultaneously accessing data and metadata in one query operation.

After a given dataset has been initially cached in the database, subsequent queries on that data experience significantly improved performance because the data need not be loaded into the relational tables. There is no need to clear data from tables when a query has completed, though the option is available if the deployment environment indicates such an approach. If and when database space does become scarce, the oldest and least frequently referenced tables are removed to make room for newly requested data.

## **3. Case Studies**

### **3.1 PASTA**

The LTER Network Information System is now developing both short and long-term strategies to enable a wide range of synthesis research within the LTER Network and the broader scientific community. One such strategy is the Provenance Aware Synthesis Tracking Architecture (PASTA) framework (Servilla et al. 2006). This modular framework is designed to support automated extraction and loading of site-based data into a permanent and persistent archive, which can be used as a data resource for synthesis research.

A foundation module of the PASTA framework, called the "Parser/Loader", is based on the Data Manager library specification. There are over 6,000 EML documents available in community Metacat servers that have been contributed by LTER as of early 2008. Most of these EML documents describe tabular data that are easily accessible by functions of the Data Manager library. It is the goal of the Parser/Loader module to automatically update PASTA's archive content when new versions of data are available at each participating LTER site. Components of the PASTA framework are now in use by the EcoTrends Web Portal (<http://www.EcoTrends.info>), a collaborative project (Peters and Laney 2006, Laney and Peters 2006) between the LTER Network and other local, state, and federal agencies and institutions to promote the use of long-term ecosystem data for synthesis research.

### **3.2 SBC-LTER**

At Santa Barbara Coastal LTER, the Data Manager library is employed as part of a web application written in Ruby that allows users to query EML-described data stored in Metacat. By integrating the Data Manager within the application's architecture, SBC has eschewed custom relational database design and embraced a flexible solution for providing robust data access features to LTER data.

The particular SBC datasets that motivated development of the query interface contain spatially and temporally specific records. Because this data has potentially high observation frequency, one of the overarching goals was to provide a mechanism for users – be they

scientists, students or other informed parties – to limit the number of records retrieved based on date, time and location parameters. Queries may further restrict results to specific data columns before the output is finally delivered in a zipped CSV file format.

### **3.3 FIRST**

A large component of the Faculty Institutes for Reforming Science Teaching (FIRST) project involves the extension of existing EML, Data Manager, Metacat and Morpho technologies to enable data analysis in the domain of science education. The expectation is that tools for both capturing and querying educational assessment data will be born from this evolution and provided to professors, instructors and researchers alike.

By incorporating the eml-dataset module in the design of the new FIRST metadata schema, the powerful features included in the Data Manager can be exploited to quickly provide query capabilities without the substantial overhead of creating a custom relational database solution. The small number and relatively simple structure of data files necessary to fully capture the FIRST assessment response data make this project an ideal proofing ground for incorporating Data Manager as a valuable tool for mediating between data and metadata.

### **4. Difficulties**

Users of the Data Manager API will find that it operates optimally when processing perfectly or near-perfectly annotated Datapackages. This reliance on human entered metadata frequently compromises the utility of the tool. Without complete data descriptions, it becomes difficult or impossible to intuit reasonable table schema in which to house the data. Moreover, the actual data must to be relatively “clean” in that data types match between metadata and data.

As the Data Manager becomes more ubiquitous among analysts, we should see a pattern of metadata and data quality improvement develop throughout the KNB. Early adopters might be frustrated by the scarcity of well-described data in the KNB and could react by seeking alternate, roll-your-own solutions for acquiring and querying data. If the community can increasingly build a reliance on and a demand for Data Manager features that require accurate and complete metadata, then quality metadata will follow.

### **5. Future roadmap**

Rather than be limited to invoking Data Manager via Java code, a specialized XML syntax for specifying queries and returning results could free clients from some of the burdens inherent in using Java. We foresee this feature being further integrated into a Metacat deployment, such that locating and querying heterogeneous data would be a standard service and would require no direct client-side use of the Data Manager.

While the Data Manager does well to expose data for researchers, it has some semantic limitations. Intimate knowledge of the data is still required in order to effectively retrieve meaningful results especially when joining tables. As EML evolves in concert with new approaches to using ontologies to capture data relationship semantics (Madin et al. 2007, Madin et al. 2008), we will likely find increased support for “intelligent” Datapackages that are able to more efficiently produce synthetic results for scientists.

The Data Manager library will ultimately reward researchers who invest in metadata entry and data annotation. Their prize will be high-quality, meaningful data sets that are available with minimal overhead.

## Acknowledgements

This material is based upon work supported by The National Science Foundation in collaboration with NCEAS (UC Santa Barbara), University of New Mexico (Long Term Ecological Research Network Office), San Diego Supercomputer Center, University of Kansas (Center for Biodiversity Research), University of Vermont, University of North Carolina, Napier University, Arizona State University, and UC Davis.

## References

- Andelman, S.J., C.M. Bowles, M.R. Willig, and R.B. Waide. 2004. Understanding environmental complexity through a distributed knowledge network. *BioScience* 54:240–246.
- Fegraus, E.H., S. Andelman M.B. Jones, and M. Schildhauer. 2005. Maximizing the value of ecological data with structured metadata: An introduction to Ecological Metadata Language (EML) and principles for metadata creation. *Bull. Ecol. Soc. Amer.* 86:158–168.
- Jones M B, Schildhauer M, Reichman O J, and Bowers S. 2006. The new bioinformatics: integrating ecological data from the gene to the biosphere. *Annual Review of Ecology, Evolution, and Systematics.* 2006. 37:519–544.
- Laney, C.M. and D.P.C. Peters, 2006. EcoTrends in Long-Term Ecological Data: a collaborative synthesis project, introduction and update. *ILTER DataBits*, Spring 2006, (<http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/06spring/>)
- Ludäscher B., Altintas I., Berkley C., Higgins D., Jaeger-Frank E., Jones M., Lee E., Tao J., Zhao Y. 2006. Scientific Workflow Management and the Kepler System. Special Issue: Workflow in Grid Systems. *Concurrency and Computation: Practice & Experience* 18(10): 1039-1065.
- Madin J. S., Bowers S., Schildhauer M., and Jones M. B. 2008. Advancing ecological research with ontologies. *Trends in Ecology and Evolution* 23 (3): 159-168. doi:10.1016/j.tree.2007.11.007
- Madin, J. S., S. Bowers, M. Schildhauer, S. Krivov, D. Pennington and F. Villa. 2007. An ontology for describing and synthesizing ecological observational data. *Ecological Informatics*, 2 (3): 279-296. doi:10.1016/j.ecoinf.2007.05.004
- O'Brien, M. and C. Burt, 2007. A Query Interface for EML dataTables *ILTER DataBits*, Spring 2007, (<http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/07spring/>)
- Peters, D.P.C. and C.M. Laney, 2006. EcoTrends in long-term ecological research project. *Jornada Trails* Vol. 10, Issue. 1, (<http://jornada-www.nmsu.edu/site/pubs/newsletr/jornv10i1.pdf>)
- Servilla, M., J. Brunt, I. San Gil, and D. Costa, 2006. PASTA: A Network-level Architecture Design for Generating Synthetic Data Products in the LTER Network. *ILTER DataBits*, Fall 2006, (<http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/06fall/>).



## VIRTUAL SENSOR-POWERED SPATIOTEMPORAL AGGREGATION AND TRANSFORMATION: A CASE STUDY ANALYZING NEAR-REAL-TIME NEXRAD AND PRECIPITATION GAGE DATA IN A DIGITAL WATERSHED

Yong Liu<sup>†</sup>, David J. Hill<sup>†</sup>, Tarek Abdelzاهر<sup>‡</sup>, Jin Heo<sup>‡</sup>, Jaesik Choi<sup>‡</sup>, Barbara Minsker<sup>§</sup>,  
David Fazio\*

<sup>†</sup>National Center for Supercomputing Applications (NCSA) 1205 W. Clark St. Urbana, IL 61801, USA. <sup>‡</sup>Department of Computer Science, University of Illinois at Urbana-Champaign 201 N Goodwin Ave, Urbana, IL 61801, USA. <sup>§</sup>NCSA and Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign 205 North Mathews Ave Urbana, IL 61801, USA. \*U.S. Geological Survey, Illinois Water Science Center 1201 W. University Ave. Suite 100 Urbana, Illinois 61801, USA

### Abstract

In this paper, we describe the case for creating and managing “virtual sensors” for near-real-time sensor data aggregation and transformation with a case study in a watershed near Chicago. We explore various levels of abstractions of “virtual sensors”, and how the virtual sensor concept and digital watershed tool can help facilitate community participation and build consensus on using and re-purposing the near-real-time data. We describe our proposed approach on aggregating NEXRAD data and on-ground *in-situ* precipitation gage data in near-real-time for anomaly detection purpose.

**Keywords:** Virtual sensor, NEXRAD, precipitation, gage station, reuse, repurpose, spatiotemporal aggregation and transformation, digital watershed, sensor web

### 1. Introduction

Physical, chemical and biological sensor networks have been used and are increasingly being deployed for measuring various environmental conditions and processes. For example, the National Weather Service (NWS) Next Generation Weather Radar (NEXRAD) system has been operational since 1997 providing meteorological observations that have been used for weather forecasting [Fulton *et al.*, 1998]. Recent advances in cyberinfrastructure technologies are allowing researchers to use large quantities of heterogeneous sensor data for furthering the understanding of large-scale environmental processes, as well as in monitoring and modeling the quality of the environments in which the sensors are deployed [Liu *et al.* 2007]. In their endeavors, researchers often seek the “raw” sensor data in order to ensure that their analyses will be free from bias, but usually such data has already been processed in certain ways. Furthermore, acquiring sensor data suitable for a particular application (e.g. at an appropriate spatiotemporal resolution) requires the repurposing of sensor data which might be beyond the scope of the original sensor design and deployment. Similar to the recent revolution of computer server virtualization (or virtual machines) [Oguchi and Yamamoto, 2008], it is now necessary to consider virtualization of sensors and sensor networks so that existing deployments of sensor networks and their measurements can be easily repurposed and used in new ways. This paper will explore the levels of virtual sensor abstraction and discuss acceptance of certain data quality control methods, coordinate transformations, measurement fusion and aggregations. A case study on aggregating NEXRAD data and on-ground *in-situ* precipitation gage data in a digital

watershed is presented. Discussions on the process of validating virtual sensors through community participation and review are also presented, as are implications for national environmental observatories such as the WATERS Network.

## 2. Related Work

The concept of a “virtual sensor” has been presented by authors in diverse areas of research. For example, virtual sensors have been defined as new predictions or higher-level concepts based on applying machine learning or artificial intelligence methods to multiple original sensor signals [see, *e.g.*, Ibarguengoytia and Reyes 2006; Persson *et al.* 2007, Peñarrocha *et al.* 2006]. In the wireless sensor network domain, virtual sensor observations are based on computation or aggregation of in-network sensor measurements. Such computation is usually based on standard SQL aggregation queries such as MIN, MAX, or SUM (see, *e.g.*, Global Sensor Network (GSN) by Aberer *et al.* 2007). Virtual sensors are also used in feedback-control applications along with “virtual actuators” [Ciciriello *et al.* 2006]. A virtual sensor network is also proposed to dynamically reconfigure sensor nodes for different purposes [Jayasumana *et al.* 2006].

In this paper, we define a virtual sensor as the product of thematic, spatial, and/or temporal transformation and aggregation of raw sensor measurements. This definition is most similar to that of Kabadayi *et al.* [2006], where heterogeneous physical sensor data are abstracted through software aggregation, although their definition does not explicitly mention spatial, temporal and thematic transformations. Their paper focuses on the development of virtual sensor Application Programming Interfaces (APIs). However, Kabadayi *et al.* do not offer a discussion on the levels of acceptable error corrections and spatial-temporal scale transformations, nor do they discuss the community approach to virtual sensors, and how virtual sensors can facilitate broad community collaborations. Furthermore, their application is construction management, not environmental management and monitoring.

## 3. Exploring the Virtual Sensor Abstraction

As a practical example of virtual sensors, a paragraph on page 108 of the recently released WATERS Network Science, Education, and Design Strategy (SEDS) Draft report [WATERS Network, 2008] reads as follows: “*Signals from arrays of individual sensors and clusters of such arrays would be combined to provide higher-level information. For example, an array of soil moisture and temperature sensors might be coupled to a microclimate array to provide a **virtual soil moisture flux sensor***”. This clearly describes the relationship between a virtual sensor and a physical sensor, that is, aggregations among physical sensors.

Since environmental processes are inherently noisy, and exist in the time-space domain, creation of a virtual sensor requires one or more of the following steps: error correction and QA/QC filtering; spatiotemporal coordinate transformations; and spatiotemporal measurement aggregations. Certain aspects of these processes are sometimes referred to as the “data ingest” process in the sensor web literature, which refers to “calibrate, gap-fill, regrid process” (see *e.g.*, Balazinska *et al.*, 2007). However, the spatiotemporal measurement aggregation step of the virtual sensor is far more complex than simple interpolation or regridding, as discussed below.

By adopting the concept of virtual sensors and providing tools to create and share virtual sensors, a higher level community participation and collaboration can be achieved. A virtual sensor created by one researcher could be used by another researcher to test different hypotheses such as new data transformations (*e.g.* Battan, [1973], Smith and Krajewski [1993], and Morin *et*

*al.* [2003] discuss different transformations from measured NEXRAD radar reflectivity to rainfall rates) or to use similar data in updated models. This re-use can build acceptance of certain processing steps for new or existing sensors via the “many eyes” approach similar to that used by wikipedia<sup>10</sup> and many other “Web 2.0” applications.

### **3.1 Error correction and QA/QC filtering**

Many agencies and data providers subject sensor data to error correction and automatic quality analysis and quality control (QA/QC) methods before distributing them. For example, over 10 error corrections and QA/QC filtering steps are performed during Stage I processing of NEXRAD precipitation data [Chrisman *et al.*, 1994; Fulton *et al.*, 1993]. We consider data processed in this way to be our first level of virtual sensor data, because different research tasks will require different levels of QA/QC and error correction. For example, data that has been subjected to intensive cleaning may be preferable for numerical models, since erroneous data may cause the model to become unstable but this type of data is unsuitable for research regarding extreme events.

### **3.2 Spatiotemporal coordinate transformations**

One commonly encountered task in the creation of an environmental virtual sensor is coordinate transformation in both time and space. Spatiotemporal coordinate transformation is well documented in many international standards (see, *e.g.*, Open Geospatial Consortium Coordinate Transformation Services [Open Geospatial Consortium, 2001]; ISO 19108:2002 Temporal Schema [ISO 19108, 2002]; ISO 19107:2003 Spatial Schema [ISO 19107, 2003]). For example, NEXRAD Level II data is georeferenced using a local polar grid centered at the radar and temporally referenced using universal time (UT), while rain gages are georeferenced using the World Geodetic System 1984 (WGS-84) and temporally referenced using the local time zone. Thus aggregating these two types of data requires both a geographic and a temporal coordinate change. An additional temporal coordinate change important to environmental sensors is the transformation from the Julian calendar to the Gregorian calendar and vice-versa.

### **3.3 Spatiotemporal measurement aggregations**

Simple measurement aggregations can be easily performed to form new virtual sensor data. For example, a new virtual wind vector sensor can be created by combining wind direction and magnitude data streams.

More complicated measurement aggregations such as up- or down-scaling sensor measurements and fusing data from multiple sensors with different spatiotemporal support (*i.e.* the region in space and/or time that the measurement represents) are more difficult to perform.

Upscaling measurements refers to the process of increasing the region in space and/or time that a particular measurement represents. Because environmental processes are highly variable in space and time, the upscaled region is often larger than the scale of the variability, and thus corrections must be made to account for effects of sub-scale fluctuations at the upscaled measurement scale (*e.g.* stochastic methods [Rubin, 2003] or spatial filtering [Beckie *et al.*, 1996]). Conversely, downscaling measurements refers to reducing the region in space and/or time that the measurement represents. Again, such scaling approach is not straight-forward when the process being measured varies at spatiotemporal scales smaller than the original measurement. Finally, fusing sensor data with different spatiotemporal support requires both

---

<sup>10</sup> <http://www.wikipedia.org>

up/down-scaling and a relationship between the sensor measurements which takes into account their correlations and the expected measurement error of each sensor. For example, the Multi-sensor Precipitation Estimator (MPE) measurements (discussed in Section 4) fuse rain gage data (point spatial support, minute temporal support, produced at a frequency of minutes to days) with radar data ( $\sim 2 \text{ km}^2$  spatial support, point time support, produced every 6 to 10 minutes) and with infrared satellite data ( $\sim 16 \text{ km}^2$  spatial support, point time support, produced every 15 minutes) [Kondragunta 2007]. This fusion requires scaling of the measurements in space and time and spatiotemporal interpolation such that the measurements from the different sensors reflect congruent quantities of rainfall and a relationship between the gages, radar, and satellite measurements that takes into account their measurement accuracy.

#### 4. A Case Study

In this case study, we discuss the development of a virtual sensor for precipitation in the Salt Creek Watershed, which is located in the greater Chicago region in Illinois. In the following sections, information and issues related to NEXRAD and precipitation gages in Salt Creek are presented, followed by a detailed description on how to create a virtual precipitation sensor. Preliminary results will be available for the conference presentation.

##### 4.1 NEXRAD

The NEXRAD system is composed of approximately 160 radar sites located throughout the United States and selected overseas areas. These radar sites measure reflectivity, radial velocity and spectrum width of the radar echoes returned from volumes within the atmosphere. These volumes are defined by a polar grid centered at the radar. The Radar Product Generator (RPG) creates 41 “products” from the three measurements made by each radar via calibrated transformations and (often) threshold-based QA/QC. These products represent estimates of meteorological process variables such as hourly precipitation, tornadic vortex signature, hail index, or severe weather probability. More information about the RPG is given by Klazura and Imy [1993] and Fulton *et al.* [1998]. The data from the NEXRAD system is divided into a hierarchy that indicates the increasing amount of preprocessing, calibration, and quality control performed [Klazura and Imy, 1993; Fulton *et al.*, 1998; Wang *et al.*, 2008]. Thus, these data (except in the case of Stage I, Level II as discussed below) are all virtual sensor data by our definition.

Stage I data refers to data from a single radar site, and is further subdivided into Level II and Level III data, which refer to the original three measurements made by the radar and the 41 products generated by the RPG, respectively. Stage II provides estimates of hourly rainfall accumulations that fuse NEXRAD Level III data with rain gage measurements averaged over a 4 km by 4 km [Fulton *et al.* 1998; Seo, 1999]. Stage III refers to a mosaic of Stage II products from multiple radars that cover an entire forecasting region of a NWS River Forecast Center. Finally, Multi-sensor Precipitation Estimator (MPE) refers to hourly rainfall accumulations that fuse NEXRAD Stage III and Geostationary Operational Environmental Satellite (GOES) products [Fulton *et al.*, 2002].

The data types and transformations used to create the Level III, Stage II and III, and MPE data are tailored to the needs of the NEXRAD agencies. Thus, if a researcher was interested in shorter duration accumulations (*e.g.* 20 min.) it would be necessary to construct these data from the Level II data. Recently, Krajewski *et al.* [2008] presented the Hydro-NEXRAD prototype, a system that will provide researchers with NEXRAD Level II data at the watershed level. This

system facilitates the transformation of Level II data using a set of predefined operations to achieve a customized output. Note that, currently, Hydro-NEXRAD is not designed for near-real-time transformation and aggregation of the Level II data, nor does it allow researchers to implement their own transformations to be added to the set of predefined operations.

#### 4.2 Precipitation Gages

The United States Geological Survey (USGS) has installed and maintains several tipping bucket rain gages within the Salt Creek watershed. These gages register volumes of water falling on a .03 m<sup>2</sup> area in 0.0254 cm increments (referred to as a tip). The tip data stream from a particular gage is sent to a programmable logic controller (PLC), which records the cumulative volume in 5-minute intervals and outputs the data to a spread spectrum radio. Failed five minute transmissions are indicated with a numerical flag. The USGS automatically removes the data flagged as failed transmissions and publishes them in the USGS National Water Information System: Web Interface (NWIS-web). USGS personnel visit the sensors on a regular basis and download locally logged data that is used for *post facto* QA/QC of the telemetered data, predominantly the removal of false zeros (*i.e.* failure of the sensor to measure falling rain).

#### 4.3 A New Virtual Precipitation Sensor

We are constructing a new virtual sensor for the Salt Creek watershed that produces measurements of 20-minute rainfall accumulations at the gage locations (with the spatial support of the gages), which merges data collected by the regional NEXRAD site (call sign KLOT) and the gages maintained by the USGS. Since NEXRAD precipitation observations at this spatiotemporal scale are not created by the RPG, Level II data will be required. The virtual sensor data stream will be produced via the following workflow steps:

1. Convert the Level II reflectivity to rainfall rates using the convective Z-R relationship [Fulton *et al.* 1998]. This is the same relationship used to create the Level III hourly precipitation accumulation product for the radar.
2. Perform QA/QC on the radar data to remove observations below the signal to noise ratio, and observations that are range ambiguous. Both these types of observations are indicated by numeric flags within the Level II data.
3. Perform QA/QC on the gage data to remove failed transmissions indicated by a numeric flag in telemetered rain gage data.
4. Accumulate both the radar and gage data in time to be collocated (in time) in 20 minute intervals.
5. Map the Level II, radial local plane coordinates onto WGS-84 geodetic coordinates (used to locate the gages) to facilitate spatial interpolation.
6. Spatially interpolate the 20 minute radar rainfall estimates to collocated estimates at the gage locations.
7. Fuse the 20 minute accumulation gage and radar data using a dynamic Bayesian method as suggested by [Hill *et al.*, 2007] to produce a robust estimate of the 20 minute precipitation accumulations at the gage locations.

A prototype digital watershed tool is being built to allow us to experiment with various types of virtual sensors for this case study. At the core of the prototype digital watershed is a light-weight virtual sensor middleware (written in the python programming language), which has

the capability to perform lightweight, in-memory, near-real-time data retrieval, transformation and aggregation [which we call filters, such as converting ESRI shapefiles to KML (Keyhole Markup Language) files], metadata extraction from KML files, and interfacing with more computationally intensive workflow-based tasks. We have chosen KML because it has recently been approved as one of the OGC standards [OGC KML, 2008]. A Google Map-based web user interface can show different KML files as different data layers and also allow users to contribute data by submitting new KML files. The idea of using KML is similar to the one proposed in ObsKML [2008]. From the end user perspectives, each KML file could represent a new virtual sensor (although we should exclude those non-changeable geographical objects such as sensor stations or watershed polygons etc.). Research communities such as WATERS Network could potentially benefit from such new virtual sensors for many environmental and hydrological applications. In addition, community users can modify the virtual sensor workflow to meet their own specific needs (such as experimenting different temporal intervals for anomaly detection or modeling) and then re-publish their own versions of virtual sensors and workflows. Virtual sensors and associated community workflows thus become sharable community resources. Such cyberinfrastructure is being developed at NCSA and will greatly improve the community collaboration and participation.

Note that data streaming middlewares such as RBNB [Tilak *et al.*, 2007] could be integrated and used to perform the data streaming task in the future, although that is beyond the scope of this paper.

## 5. Conclusion and future Work

This paper explores the usage of virtual sensors in near-real-time environmental sensor networks with a case study that proposes a virtual precipitation sensor in a participatory digital watershed near Chicago. We describe various levels of virtual sensors and their spatiotemporal transformations and how they are relevant to environmental observatories. Often on-the-fly transformations can be done by applying some light-weight filtering operations, while more computationally intensive transformations can be done through workflow systems such as the CyberIntegrator system being developed by NCSA [2008]. Virtual sensors can be considered new, near-real-time sensor data sources and thus, can be reused among community researchers. Provenance-aware virtual sensors are therefore valuable tools for community collaboration where different users can examine how the virtual sensors are derived. Provenance technologies are also being developed by NCSA and we will explore such integration with the prototype system presented in this paper.

We think the concept of virtual sensors and the virtualization of sensor networks will allow diverse user communities to access and modify sensor data and potentially even provide new virtual sensor data streams over the internet in near-real-time. This is similar to the Microsoft SenseWeb [Kansal *et al.* 2007] idea, but with an extension from physical sensors to more broadly defined virtual sensors.

Our future exploration of virtual sensors will also consider privacy filtering as another layer of transformation. The privacy issue will become evident when large-scale environmental monitoring sensor networks are coupled with citizen science-type participatory sensing [Abdelzaher *et al.*, 2007; Cuff *et al.*, 2008]. For example, in the city of Chicago, we will soon have access to near-real-time consumer water usage data through smart water meters deployed at the residential household level. This demands proper privacy protection when such data are provided for environmental and water resource study.

## Acknowledgements

The authors thank UIUC/NCSA Adaptive Environmental Sensing and Information Systems (AESIS) initiative for funding this work through Grant 1-200250-251024ICR. We would also like to acknowledge the Office of Naval Research, which partially supports this work as part of the Technology Research, Education, and Commercialization Center (TRECC) (Research Grant N00014-04-1-0437) managed by NCSA. The authors also thank Seong-Gon Kim for helping with the prototype digital watershed implementation.

## References

- Abdelzaher, T., Y. Anokwa, P. Boda, J. Burke, D. Estrin, L. Guibas, A. Kansal, S. Madden, and J. Reich (2007), Mobiscopes for human spaces, *IEEE Pervasive Computing*, 6(2), 20-29, doi: 10.1109/MPRV.2007.38.
- Aberer, K., Hauswirth, M., Salehi, A. (2007), Infrastructure for data processing in large-scale interconnected sensor networks, *Mobile Data Management (MDM)*, Germany, 2007. Available at: <http://lsirpeople.epfl.ch/salehi/papers/GSN-MDM07.pdf>
- Balazinska, M., A. Deshpande, M. J. Franklin, P. B. Gibbons, J. Gray, S. Nath, M. Hansen, M. Liebhold, A. Szalay, and V. Tao (2007), Data management in the worldwide sensor web, *IEEE Pervasive Computing*, 6(2), 30-40, doi: 10.1109/MPRV.2007.27.
- Battan, L. J. (1973), *Radar Observations of the Atmosphere*, Chicago: The University of Chicago Press.
- Beckie, R. Aldama, A. A., and Wood, E. F. (1996), Modeling the large scale dynamics of saturated groundwater flow using spatial-filtering theory: 1. Theoretical development, *Water Resources Research*, 32(5), 169-1280
- Chrisman, J., Rinderknecht, D., and Hamilton, R. (1994), WSR-88D clutter suppression and its impact on meteorological data interpretation. Preprints, First WSR-88D User's Conference, Norman, OK, WSR-88D Operational Support Facility, 9-20.
- Ciciriello, P., L. Mottola, and G. P. Picco (2006), Building virtual sensors and actuators over logical neighborhoods, *International Workshop on Middleware for Sensor Networks, MidSens 2006*. Co-located with *Middleware 2006*, Melbourne, 28 November 2006 through 28 November 2006.
- Cuff, D., M. Hansen, and J. Kang (2008), Urban sensing: Out of the woods, *Communications of the ACM*, 51(3), 24-33, doi: 10.1145/1325555.1325562.
- Droegemeir, K., Kelleher, K., Crum, T., Levit, J.J., Del Greco, S.A., Miller, L., Sinclair, C., Benner, M., Fulker, D.W., and Edmon, H. 2002. Project CRAFT: A testbed for demonstrating the real time acquisition and archival of WSR-88D Level II Data. 18th Int. Conf. on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology., 13-17 January, American Meteorological Society, Orlando, Florida, 136-139.
- Fulton, R. A. (1998), WSR-88D polar-to-HRAP mapping, Technical Memorandum, Hydrologic Research Laboratory, Office of Hydrology, National Weather Service, Silver Spring, Maryland. Available at: <http://www.weather.gov/ohd/hrl/papers/wsr88d/hrapmap.pdf>
- Fulton, R. A., Breidenbach, J. P., Seo, D.-J., Miller, D. A., and O'Bannon, T. (1998), The WSR-88D rainfall algorithm, *Weather and Forecasting*, (June 1998), 377-395.
- Fulton, R.A. (2002), Activities to improve WSR-88D radar rainfall estimation in the National Weather Service, *Proceedings of the Second Federal Interagency Hydrologic Modeling Conference*, Las Vegas, Nevada, July 28-August 1, 2002

- Hill, D.J., Minsker, B.S., and Amir, E. (2007), Real-time Bayesian anomaly detection for environmental sensor data, Proceedings of the 32nd Congress of IAHR, International Association of Hydraulic Engineering and Research, Venice, Italy.
- Ibarguengoytia, P. H., and A. Reyes (2006), Constructing virtual sensors using probabilistic reasoning, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 4293 LNAI, 218-226.
- ISO 19107 (2003): Geographic information -- Spatial schema. Available at:  
[http://www.iso.org/iso/iso\\_catalogue/catalogue\\_tc/catalogue\\_detail.htm?csnumber=26012](http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=26012)
- ISO 19108 (2002): Geographic information -- Temporal schema (2002). Available at:  
[http://www.iso.org/iso/iso\\_catalogue/catalogue\\_tc/catalogue\\_detail.htm?csnumber=26013](http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=26013)
- Jayasumana, A. P., H. Qi, and T. H. Illangasekare (2007), Virtual sensor networks - A resource efficient approach for concurrent applications, 4th International Conference on Information Technology-New Generations, ITNG 2007, Las Vegas, NV, 2 April 2007 through 4 April 2007.
- Kabadayi, S., A. Pridgen, and C. Julien (2006), Virtual sensors: Abstracting data from physical sensors, WoWMoM 2006: 2006 International Symposium on a World of Wireless, Mobile and Multimedia Networks, Buffalo-Niagara Falls, NY, 26 June 2006 through 29 June 2006.
- Kansal, A., Suman Nath, Jie Liu, and Feng Zhao, (2007) "SenseWeb: An Infrastructure for Shared Sensing," IEEE Multimedia. Vol. 14, No. 4, pp. 8-13, October-December 2007.
- Klazura, G. E. and Imy, D. A. (1993), A description of the initial set of analysis products available from the NEXRAD WSR-88D system, Bulletin of the American Meteorological Society, 1293-1311.
- Kondragunta, C. (2007), NOAA's Water Resource Information, Presented to Space Policy Institute, George Washington University, Available at  
[http://www.gwu.edu/~spi/Chandra\\_Kondragunta-NOAA's%20Water%20Resources%20Information.pdf](http://www.gwu.edu/~spi/Chandra_Kondragunta-NOAA's%20Water%20Resources%20Information.pdf)
- Krajewski, W. F., Kruger, A., Smith, J. A., Lawrence, R., Goska, R., Domaszczynski, P., Gunyon, C., Seo, B. C., Baek, M. L., Bradley, A. A., Ramamurthy, M. K., Weber, W. J., Delgreco, S. A., Nelson, B., Ansari, S., Murthy, M., Dhutia, D., Steiner, M., Ntelekos, A. A., Villarini, G. (2008) Hydro-NEXRAD: community resource for use of radar-rainfall data, CUAHSI CyberSeminar April 25, 2008. Available at  
<http://www.cuahsi.org/cyberseminars/Krajewski-20080425.pdf>
- Liu, Y., Minsker, B., Hill, D. (2007), Cyberinfrastructure Technologies To Support QA/QC and Event-Driven Analysis Of Distributed Sensing Data, International Workshop on Advances in Hydroinformatics, 4 - 7 June 2007, Niagara Falls Canada
- Morin, E., Krajewski, W. F., Goodrich, D. C., Gao, X., and Sorooshian, S. (2003), Estimating rainfall intensities from weather radar data: The scale-dependency problem, Journal of Hydrology, 4, 782-797.
- NCSA TRECC year-8 Project Executive Summary (2008), Available at:  
<http://cet.ncsa.uiuc.edu/GSoC/TRECC-EXEC.pdf>
- NOAA News. (2004), <http://www.noaanews.noaa.gov/stories2004/s2208.htm> . Accessed 04/11/2008.



- ObsKML (2008), [http://nautilus.baruch.sc.edu/twiki\\_dmcc/bin/view/Main/ObsKML](http://nautilus.baruch.sc.edu/twiki_dmcc/bin/view/Main/ObsKML), Accessed 04/14/2008
- Oguchi, Y., and T. Yamamoto (2008), Server virtualization technology and its latest trends, *Fujitsu Scientific and Technical Journal*, 44(1), 46-52.
- Open Geospatial Consortium, (2001), Coordinate Transformation Services Available at: <http://www.opengeospatial.org/standards/ct>
- OGC KML (2008), OGC® Approves KML as Open Standard, OGC Press release, April 14, 2008. Available at: <http://www.opengeospatial.org/pressroom/pressreleases/857>
- Peñarrocha, I., R. Sanchis, and P. Albertos (2006), Design of low cost virtual sensors, IMTC'06 - IEEE Instrumentation and Measurement Technology Conference, Sorrento, Italy, 24 April 2006 through 27 April 2006.
- Persson, M., T. Duckett, C. Valgren, and A. Lilienthal (2007), Probabilistic semantic mapping with a virtual sensor for building/nature detection, 2007 IEEE International Symposium on Computational Intelligence in Robotics and Automation, CIRA 2007, Jacksonville, FL, 20 June 2007 through 23 June 2007.
- Rubin, Y. 2003. Applied Stochastic Hydrogeology. Oxford University Press, New York.
- Seo D. J. , Breidenbach J. P. , Johnson E. R. (1999), Real-time estimation of mean field bias in radar rainfall data, *Journal Hydrology*, 223, 131-47.
- Smith, R. E. and Krajewski, W. F. (1993), A modeling study of rainfall rate-reflectivity relationships, *Water Resources Research*, 29, 2505-2514.
- Tilak, S. P. Hubbard, M. Miller, T. Fountain (2007), The Ring Buffer Network Bus (RBNB) DataTurbine Streaming Data Middleware for Environmental Observing Systems, e-Science, 10/12/2007, Bangalore, India
- Wang, X., Xie, H., Sharif, H., and Zeitler, J. (2008), Validating NEXRAD MPE and Stage III precipitation products for uniform rainfall on the Upper Guadalupe River Basin of the Texas Hill Country, *Journal of Hydrology*, 348, 73-86.
- WATERS Network SEDS draft report (2008). Science, Education, and Design Strategy for the WATER and Environmental Research Systems Network (SEDS) - February 27, 2008. Available at <http://www.watersnet.org/docs/SEDS-20080227-draft.pdf>

## **DATA COLLABORATION FOR LARGE-SCALE REGIONAL SURVEYS IN SOUTHERN CALIFORNIA**

**Shelly L. Moore†, Shelly Walther‡, and Larry D. Cooper†**

† Southern California Coastal Water Research Project, 3535 Harbor Blvd. Suite 110, Costa Mesa, CA 92626 ‡ Sanitation Districts of Los Angeles County, 1955 Workman Mill Road, Whittier, CA 90601

### **Abstract**

Southern California has some of the most visited beaches and recreational waters in the world and a major challenge is ensuring environmental and human health in these areas. Monitoring of conditions often involves independent agencies studying only a small area relative to the Southern California region as a whole. In 1994 a pilot project was initiated to conduct a regional survey of the Southern California Bight (from Point Conception, CA to the U.S./Mexico border) to assess environmental conditions on a large scale. A major challenge in this survey and those that followed in 1998 and 2003 were the large number of participants, who typically use different methods for data collection and storage. The information management system used in these three studies stressed collaboration and teamwork among all of the participating agencies and programs. This paper provides a summary of the collaborative process and success of these surveys from an information management perspective.

**Keywords:** Information Management, Collaboration, Environmental, California

### **1. Introduction**

Southern California beaches and marine waters provide unique year-round recreational opportunities and are some of the most visited sites in the world. With an estimated 150 million visitors annually to its beaches and recreational waters, southern California provides a large amount (over \$9B) of money to the local economy (Schiff *et al.* 1999). Monitoring of these waters costs and estimated \$31M (Schiff *et al.* 2002) and is an important part of ensuring human and environmental health. Much of the monitoring is typically done by dischargers and local agencies as part of their permit requirements assigned by state and federal government agencies. This monitoring is often limited to small areas and the data is usually not available or in the proper format to use on a larger scale. In 1994 a unique regional monitoring pilot project program was initiated to make environmental health assessments for the Southern California Bight (SCB; Cross and Weisberg 1996), an area of the mainland shelf from Point Conception, California to the U.S./Mexico border. This program was the first of its kind on the west coast and its success led to similar regional surveys in 1998 (Bight 98) and 2003 (Bight 03). Over time these projects not only grew in scope, but also in the number of participants, from twelve agencies in 1994 to over sixty agencies in both 1998 and 2003.

The large number of participants in these surveys made for huge challenges in information management. The first major obstacle was to ensure collaboration among all participants to standardize methods for data collection and reporting. This was made even more difficult given that the data types to be collected included a wide range of disciplines, such as sediment chemistry, sediment toxicity, benthic infauna, trawl caught fish and invertebrates, and

water quality. Another major challenge was in collating the data into a single unified data system that made them useable for easy data analysis and reporting.

Beginning with the 1994 Pilot Project, an Information Management System (IMS) was put in place to promote an environment of collaboration and teamwork. This involved the formation of an Information Management Committee (IMC) to oversee data structure and reporting requirements. Many of the agencies and programs participating in the survey sent representatives to the IMC and were made part of the data management process, thereby developing data management structures using a consensus based approach. The success of this type of collaboration has led to subsequent surveys in 1998 and 2003. The goal of this paper is to provide a summary of the collaborative process that led to the successful data management over these three regional surveys.

## 2. Methods and Results

The large scope of these surveys provided a unique challenge to organizing and structuring the necessary committees to ensure their success. The overall structure of the surveys varied from survey to survey depending on the focus; however, the general structure of the groups stayed the same throughout the surveys (Figure 1).

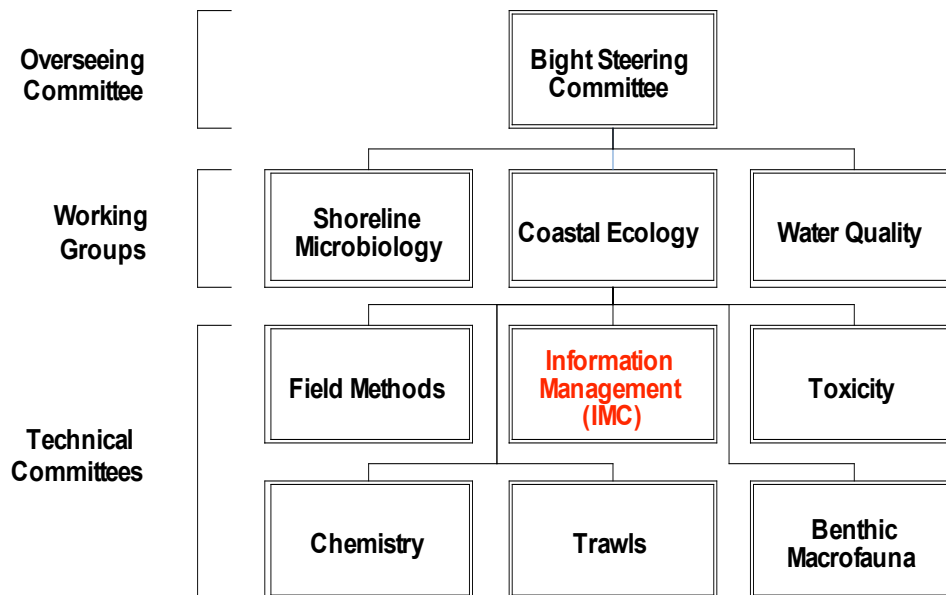


Figure 1. General structure and organization of the Bight surveys (from Bight 03).

The overall function of the Bight Steering Committee was to determine the direction and goals of each survey. In addition, they were also responsible for creating working groups and technical committees that were appropriate for meeting the needs of each survey. The primary function of the technical subcommittees was to ensure the standardization of field and lab methods. The IMC played an important role in coordinating data collection and reporting for all of the groups involved. It was crucial for the IMC to be involved with the rest of the committees to ensure an understanding of the data that was to be collected and the information that was required to make assessments.

Any agency that was involved with a survey was encouraged to send representatives to the IMC and any other technical committees. Members of each committee then selected a chairperson to oversee committee meetings and report to the Bight Steering Committee. For the IMC, the chairperson also acted as the Information Management Officer, who was responsible for coordinating the submission of data. The IMC began meeting well before any of the surveys took place, and members of this committee often consisted of information management personnel from the participating agencies. These personnel, or Agency Information Managers (AIM), would also be responsible for submitting any data generated by their agency or program. The AIM provided a single point of contact for each agency, making communication with the IMO much easier.

The IMC, using a consensus-based decision process, created an Information Management Plan (Cooper *et al.* 2003) that included all of the requirements for data tables and guidelines for data submission. A centralized database model with a relational database structure was developed at a level that made the data easy to use for project scientists (Figure 2).

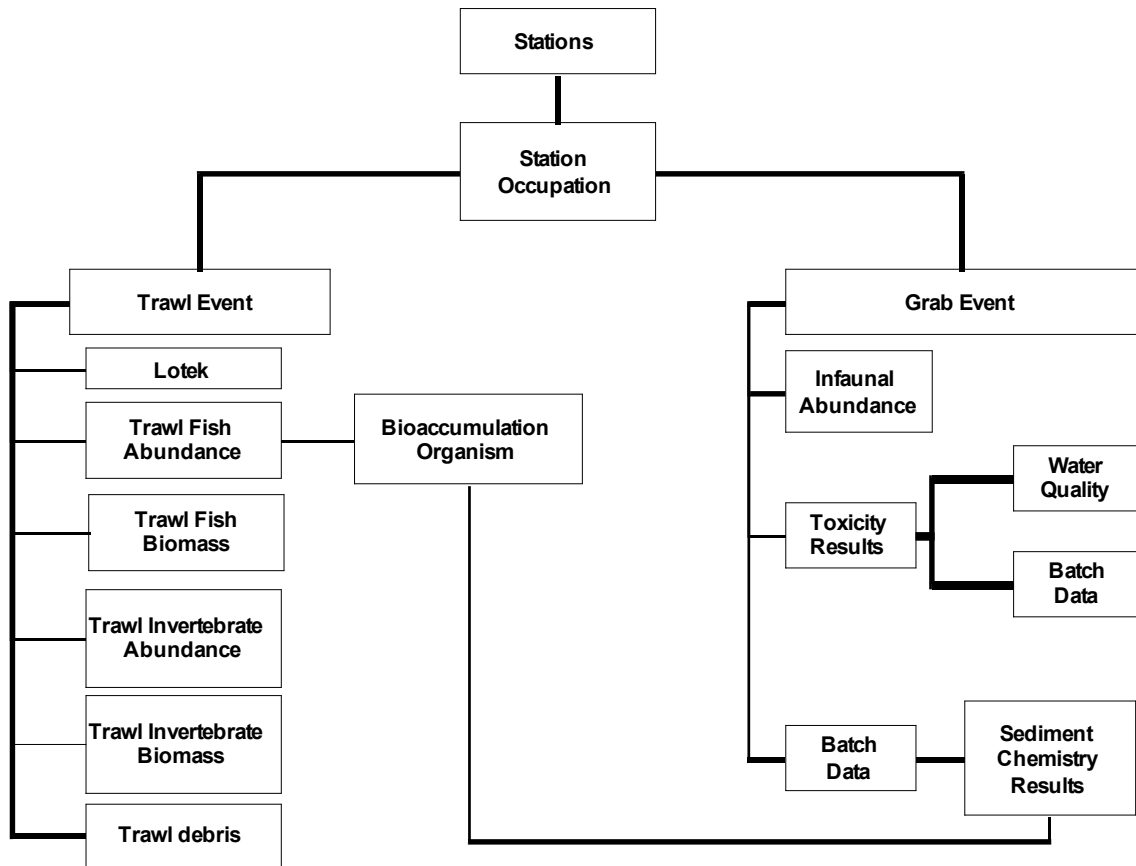


Figure 2. An example of the relational database tables used for the Coastal Ecology portion of the Southern California Bight regional surveys (from Bight 03).

Standardized Data Transfer Protocols (SDTPs) were developed for data submissions to ensure data consistency and comparability. The SDTPs allowed each agency to store data in the manner of their choice, but ensured that the data was formatted to the project standards. The SDTPs detailed the information to be submitted by each agency and included information such as field names, data types and length, and a brief description of what each field represented.

Lookup lists were created to ensure field values were consistent with the data being collected and also to ensure data quality. In addition, a Field Data System (Microsoft Access) was created to allow sampling crews to record data on field conditions directly into a database while out in the field.

The IMC developed a formalized data submission process that also included different levels of QA/QC. This process was initiated with the data collector and ended with the technical committee that was responsible for producing a report from the data (Figure 3).

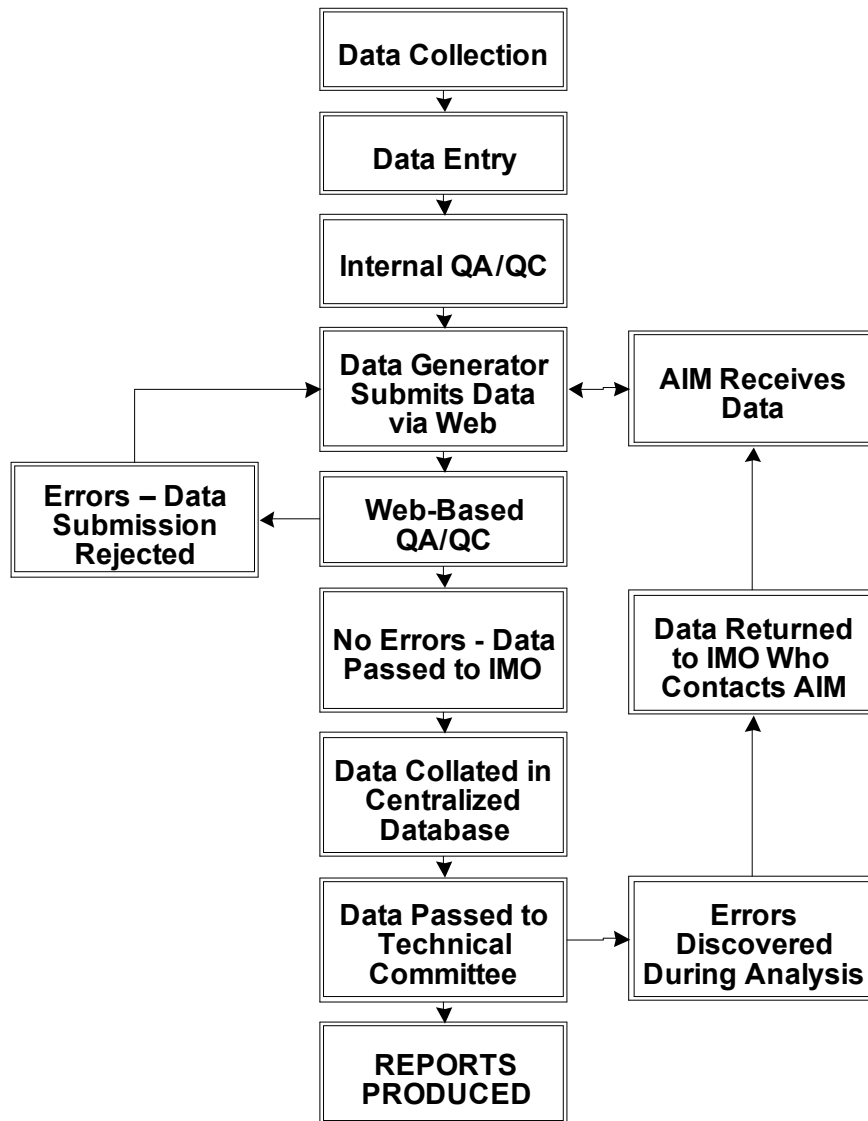


Figure 3. Simplified diagram of data submission and QA/QC process. (IMO = Information Management Officer; AIM = Agency Information Manager)

For the 1994 Pilot Project data was submitted in various file formats and each submission checked by hand for consistency and errors. As the surveys progressed, data checking became an automated process and with the Bight 03 survey, all data submissions were made via the internet (using PHP and SQL Server 2005) and data QA/QC checks were automatically processed. This

method saved both time and money for the data submitters as well as the IMO. Data submitters received instant feedback on the quality of their data and were able to fix and resubmit the data in a relatively short time. Once the data was successfully submitted it was collated and turned over to the respective technical committees for data analysis. Often errors were found during the analysis phase and brought to the attention of the IMO. Changes to the data were not made until the data generator was notified and approved of the change. Once the final survey reports were produced the data (with corresponding metadata) was then released for use by outside scientists, environmental managers and the general public.

### **3. Discussion**

The Bight surveys have provided regional assessments of the health and condition of the marine environment in southern California. These assessments assist environmental managers in making decisions by providing them information on their areas of concern in relation to all of southern California. The agencies that have participated in these surveys have developed strong working relationships and a desire to share data. The system of teamwork and consensus-based decision making was instrumental in making these surveys successful.

With each new regional survey additional lessons have been learned. The creation of the Bight Steering Committee to provide focus to the IMC and other committees has provided the necessary organizational structure to ensure the success of each survey. In addition, the information management has progressed and improved as well. Participation by information management personnel from each agency led to the successful completion of an Information Management Plan, which stipulated all of the requirements for data structure and submission. The incorporation of a web-based data submission system, for the most recent survey, not only increased the data quality but also decreased the amount of time between data collection and analysis, allowing reports to be produced in a timelier manner. Having a single point of contact with an AIM from each participant was also beneficial and decreased the need for additional communication by the IMO.

The main goal behind information management for these surveys was to share data in a common format, regardless of the data management practices of the participants. The specific methods and software programs used to accomplish this were chosen as part of the consensus-based decision process, which focused on ease of use and commonality among programs. Another key element to all of these surveys is the provision of data, after completion of the project, to outside scientists and the public. To accommodate this process, data is typically made available as simple ASCII files, downloadable via FTP, and usable in most spreadsheet or database software programs. In addition, comprehensive metadata are available that includes quality assurance classifications of the data as well as documentation of the methodologies by which the data were collected.

The next survey, scheduled for this year, will provide additional challenges as this will be the largest, most comprehensive survey yet. In addition, new agencies will be participating and new data types collected, making the need for collaboration and teamwork even greater. Information management will continue to grow in scope and change with the focus of each survey and will remain an instrumental part of their continued success.

### **Acknowledgements**

The authors wish to thank the Bight Steering Committees and all of the agencies (over sixty) that have participated in the Bight surveys. We would also like to thank all of the

individuals that have participated in the Bight Information Management Committees. All documents and reports relative to the Bight surveys are available at [www.sccwrp.org](http://www.sccwrp.org).

### References

- Cooper, L. , S. Walther, S. Moore, B. Bealer, D. Montagne, D. O'Donohue, D. Olson, H. Nguyen, K. Wisenbaker, K. Yamamoto, S. Watts, S. Meyer, L. King, M. Mengel, S. Johnson, T. Pira and M. Hoxsey. 2003. Bight'03 Information Management Plan. Southern California Coastal Water Research Project. Westminster, CA.
- Cross, J.C. and S.B. Weisberg. 1996. The Southern California Bight Pilot Project: An overview. pp. 104-108 *in*: M.J. Allen (ed.), Southern California Coastal Water Research Project 1994-95 Annual Report. Westminster, CA.
- Schiff K., J. Dorsey, and S. Weisberg. 1999. Marine microbiological monitoring in the southern California Bight. Pp 179-186. *In*: S. Weisberg and D. Hallock (eds.), Southern California Coastal Water Research Project Annual Report 1997-1998. Southern California Coastal Water Research Project, Westminster, CA.
- Schiff, K.C., S.B. Weisberg and V.E. Raco-Rands. 2002. Inventory of ocean monitoring in the Southern California Bight. *Environmental Management* 29:871-876.

## PROCESSING AND QUALITY CONTROL OF KELP FOREST COMMUNITY SURVEY DATA

**Margaret O'Brien and Shannon Harrer**

Santa Barbara Coastal LTER, Marine Science Institute, University of California, Santa Barbara, CA, USA

### **Abstract**

The Santa Barbara Coastal LTER maintains and regularly publishes many community survey data products on kelp forest biota. The logistics associated with their collection require that researchers take advantage of volunteer labor. High-quality data in time-series studies demand consistency of identification and survey methods although communication between team members is nearly impossible underwater. The Santa Barbara Coastal LTER has developed methods to train and monitor undergraduate volunteers during both data collection and processing, which combines commercial desktop spreadsheets and statistical analysis software. The system takes advantage of software features that best suit the project's needs and assure continuity. The final data product is easily integrated with scripts for publication in local and network data catalogs.

**Keywords:** quality control, SAS<sup>®</sup>, community survey, kelp forest, subtidal, Ecological Metadata Language

### **1. Introduction**

The purpose of subtidal community surveys in the Santa Barbara Coastal Long Term Ecological Research Project (SBC LTER) is to follow changes in the species composition and abundance of kelp forest biota over the long term in response to environmental change. This entails the identification and quantification of over 150 species of marine algae, invertebrates and fish using established protocols that require expertise in scientific diving. The collection of most subtidal data is personnel intensive and typically includes undergraduate student trainees, a situation that is both cost effective and provides enormous educational opportunities. The datasets resulting from these surveys are used by SBC investigators in their own research projects, and are also contributed to the LTER Network data repository for wider use. These varied uses require that data are consistently reliable, and that all processing steps are well documented.

During all stages of data collection and processing, the SBC LTER strives to integrate information management with the project's research objectives. To accomplish this goal data products are outlined jointly by research, technical, and information management staff based on expected use and scientific need. Integration is further enhanced by the use of scripted processing where data publication can be included if possible. Our processing decisions take advantage of software features that best suit our needs and assure continuity. This paper describes the procedure that is used to maintain high quality in SBC-LTER's long-term data sets pertaining to kelp forest community dynamics, while simultaneously advancing the university's and the LTER network's goal of training and education at the undergraduate level. Although underwater surveys may require special considerations, most of the procedures described here could be generally applied to other types of community surveys.



## **2. Methods**

### **2.1 Field collection**

The benefits of a quality control process in the field are clear but difficult to quantify. Subtidal community surveys are logistically complex and require a great deal of coordination, particularly since extended periods under water require the use of Self Contained Underwater Breathing Apparatus (SCUBA) and its associated safety precautions. Since extensive use is made of undergraduate trainees, a field team is generally composed of 1 or 2 expert divers and 2 or 3 undergraduate or novice divers. An expert is one who has logged hundreds of scientific dives and has extensive training and experience with the identification, natural history and monitoring of local kelp forest biota. Novice divers may be experienced in recreational diving but have typically logged few scientific dives and have little expertise in the identification and sampling of kelp forest biota.

Environmental conditions in kelp forests off Santa Barbara are usually challenging: moderate to strong wave surge and cold water temperature are the norm, visibility is usually poor, there is no audio communication, and dives are time-limited. Consequently there is little opportunity for information exchange during the survey itself, despite the obvious need for communication between experts and novice divers. Therefore, established sampling protocols are reinforced during pre-dive briefings. Divers record observations using pencil and plastic water proof paper mounted to an acrylic slate. Following each dive, data sheets are cross-checked for completeness, and the expert's knowledge of local biota is used to reinforce correct species identification and counts. The importance of these communication steps cannot be understated, and they would be required regardless of the data entry method (e.g., using electronic entry instead of paper datasheets).

### **2.2 Data processing**

The software chosen for subtidal data processing takes advantage of features that best suit a specific need: flexibility, consideration of the expertise of both creators and users, ease of data exchange, and long-term preservation. For data entry, templates are created using MS-Excel<sup>®</sup>. These are created by lab personnel, are easy to develop and customize, and can mirror the data sheets used in the field to eliminate confusion. Creators take advantage of cell validation features to control content and file server permission settings that allow templates to be edited only by certain individuals. Statistical Analysis System (SAS<sup>®</sup>) is used for all data processing. It is capable of reading the Excel<sup>®</sup> output, but all processing constants can be kept in configuration files rather than in proprietary code. All SBC data exchange is in ASCII text format, which is easily and reliably handled by SAS<sup>®</sup> structures. The flow of information from the field to data package export is outlined in Figure 1.

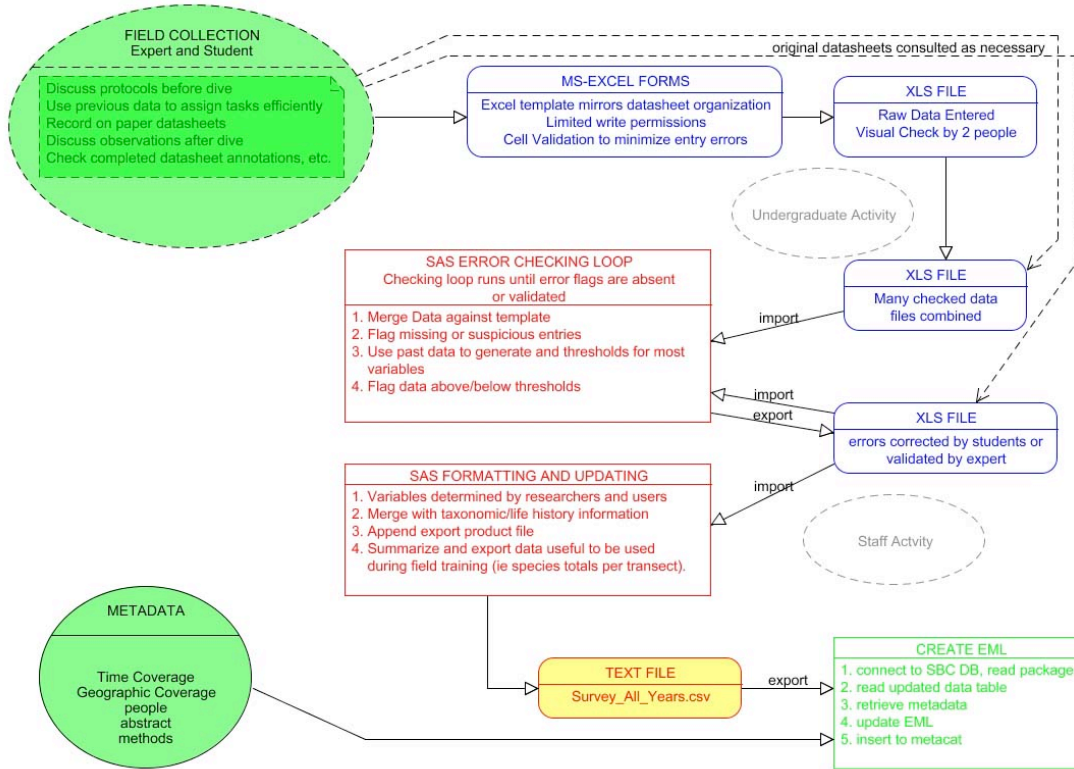


Figure 1. Data flow for subtidal surveys. Data sources are in green circles. Excel spreadsheet files are indicated in blue, SAS code steps in red.

### 2.3 MS Excel<sup>®</sup> Data Entry

Excel<sup>®</sup> data entry is performed by undergraduate trainees, and usually not by the same individual who recorded the observations. Maximum efforts are taken to minimize ambiguity and ease the transition from data on paper to electronic form. Excel column headers are identical to fields named on original data sheets and many columns are enabled with cell validation features to decrease the risk of entry errors (Fig. 2). Detailed data entry protocols inform student interns of text descriptions, and include sample electronic data sheet documents and excel screen shots detailing how information from paper data sheets should be converted into an excel data file. Once data have been entered into electronic form they are jointly checked by two interns working together: one reading the original data sheet and the other confirming that the information is correct on the Excel<sup>®</sup> spreadsheet (Fig. 1). All questions, comments and concerns are highlighted for review by research staff. After inspection and approval, research staff members combine the individual sheets into a single worksheet to be imported into SAS for processing.

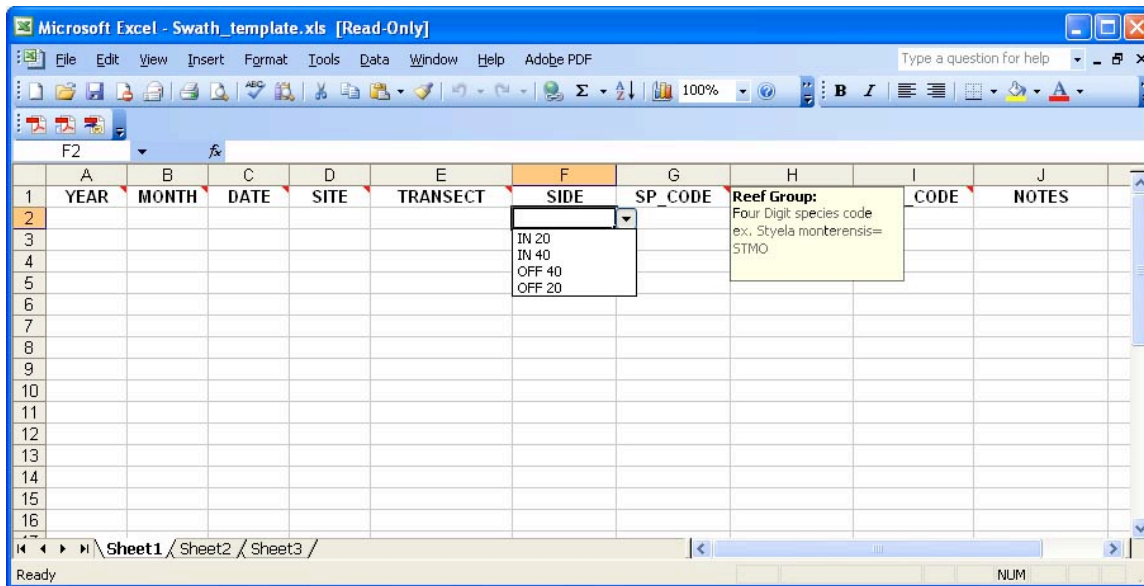


Figure 2. MS-Excel template for data entry. Content of some fields is controlled, and explanations are included in comments.

#### 2.4 Data processing using Statistical Analysis System (SAS<sup>®</sup>)

All processing is accomplished using SAS<sup>®</sup> scripts. While creating intricate checking loops can be time-consuming, for a long-term monitoring project collecting data continuously, the benefits of nearly automated processing and quality control far outweigh the costs of programming. Processing is composed of five major steps including a loop:

1. Import Excel<sup>®</sup> files into SAS<sup>®</sup> data object and merge with a template
2. Flag missing or suspicious data
3. Create and export an error summary dataset in Excel<sup>®</sup>  
[Steps 1-3 are run until data is found to be free of errors]
4. Format final data product and combine with previous data values
5. Export the updated dataset for publication

Missing or suspicious values in imported data (e.g., abundances and sizes outside of threshold values) are flagged using SAS<sup>®</sup> data steps (Fig. 3). SAS<sup>®</sup> then outputs a subset of the flagged raw data with a text description of the problematic entry. Undergraduate trainees correct all entries by referring to original data sheets and consulting with divers. An excel spreadsheet called the "SAS<sup>®</sup> Error Message Index" is used as a reference to locate and correct the problem (Fig. 4). This process is repeated until no more errors are reported. Flagged values that do not appear to be the result of an entry or observer error are left as is, presented to research staff and details of the circumstances are noted in the processing logs associated with the dataset (also a text file, Fig. 5). Lastly, data are formatted according to appropriate parameters agreed upon by the scientific, technical and information management staff, and published with EML metadata. It is important that consistent and regulated formats procedures are in place so that data can be later imported into a relational database for querying and delivery to users (O'Brien and Burt, 2007).

```

/*Merges new data with a template of sites, transects and distances
to make sure all sites transects and distances are accounted for*/
proc sort data= _raw_upc;
  by site transect distance side;
run;

data temp1;
  merge template _raw_upc;
  by site transect distance side;
run;

/*Flags the line of data if any fields are missing. Also flags sand substrate scores that have
no information for sand depth*/
data e1;
  set temp1;
  informat error $100.;
  format error $100.;
  if date= '.'d then error='missing data';
  if (date ne '.'d) and (substrte='.') then error='missing substrate';
  if (date ne '.'d) and (substrte='') then error='missing substrate';
  if (date ne '.'d) and (obs_code=.) then error='missing obs_code';
  if (date ne '.'d) and (side='') then error='missing side';
  if (date ne '.'d) and (position=.) then error='missing position';
  if (date ne '.'d) and (sand_depth=.) then error='missing sand depth';
  if (substrte='S') and (sand_depth_ le 0) then error='wrong sand depth';
  if error='' then delete;
run;
    
```

Figure 3. SAS<sup>®</sup> code fragment showing the template-merge and some error-flagging steps for missing percent cover data (UPC).

	A	B	C
1	Error Message	Survey	Description
2	Missing substrate	UPC	No data has been entered for substrate
3	Missing obs_code	UPC	No data has been entered for obs_code
4	Missing position	UPC	No data has been entered for position
5	Missing sand depth	UPC	No data has been entered for sand depth
6	Wrong sand depth	UPC	Sand depth should be greater than 0. If no sand depth is written on data sheet, ask the diver
7	Wrong substrate code	UPC	The code entered is not a known substrate code. Usually a typo or bad writing, if not ask the diver.
8	Wrong species code	UPC	The code entered is not a known species code. Usually a typo or bad writing, if in doubt ask the diver.
9	Position error	UPC	The number of species entered and the number of positions entered do not agree. Often 2 species are accidentally entered for the same position.
10	No info for quad	QUAD	Nothing data has been entered for that quad. If no species were counted enter the site/transect/date/obs_code info with a '.' for sp_code
11	Missing obs_code	QUAD	No data has been entered for obs_code
12	Missing date	QUAD	The date has not been entered
13	Wrong species code	QUAD	The code entered is not on the quad list of species code. Check for bad writing or ask the diver.
14	Missing Data	GIANT KELP	No data was entered. Data should always be entered for every section. If there is no kelp then enter all the date and site info and enter fronds=0.
15	Small Holdfast Diameter?	GIANT KELP	Holdfast measurement seems small. Make sure it is not an entry error. If the number is correct, highlight it in the data.
16	Wrong species code	SWATH	The code entered is not a known species code. Usually a typo or bad writing, if in doubt ask the diver.
17	Entry Error	SWATH	An entered code, site or transect is not a known code. Usually a typo or bad writing, if in doubt ask the diver. This could be the same as a species code error abo
18	Missing Data	SWATH	No species were recorded for this part of the transect. This is fairly uncommon and the original datasheet should be checked to make sure it is not an entry proble
19	Missing Data	ALLOMETRIC	Either an entire transect has not been entered or date/observer code was not filled in
20	Check Both Datasets	ALLOMETRIC	The number of plants in the allometric data and swath data do not add up. If there are more than 30 plants on the entire transect, 30 allometric plants should have t
21	Too many Entries	ALLOMETRIC	This is usually a case where 0s have been entered 30 times for transects with no plants. If there are no plants only 1 line of information should be entered!
22	Check Blade Data	ALLOMETRIC	Something weird is happening with one or more entries in the blade column. It may be that a '.' was entered rather than a 0. '.' should only be entered for transects
23			
24			
25			

Figure 4. The SAS<sup>®</sup> Error Message Index catalogs SAS<sup>®</sup> output error flags for students by survey. Descriptions afford information about the type of error found and the actions that should be taken to rectify the problem in MS Excel.

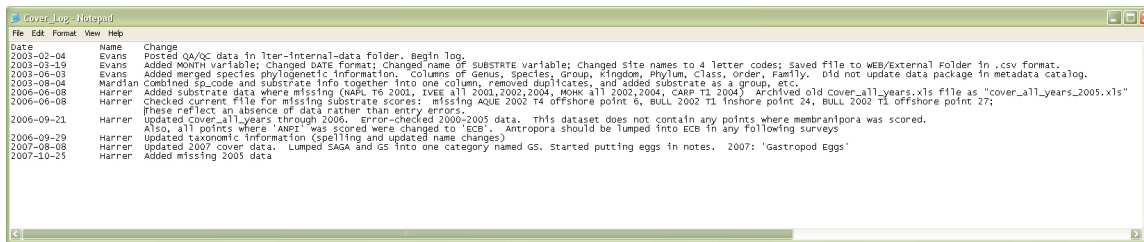


Figure 5. Sample dataset log for uniform point contact data (UPC). This file records processing events and the locations and descriptions of questionable data values.

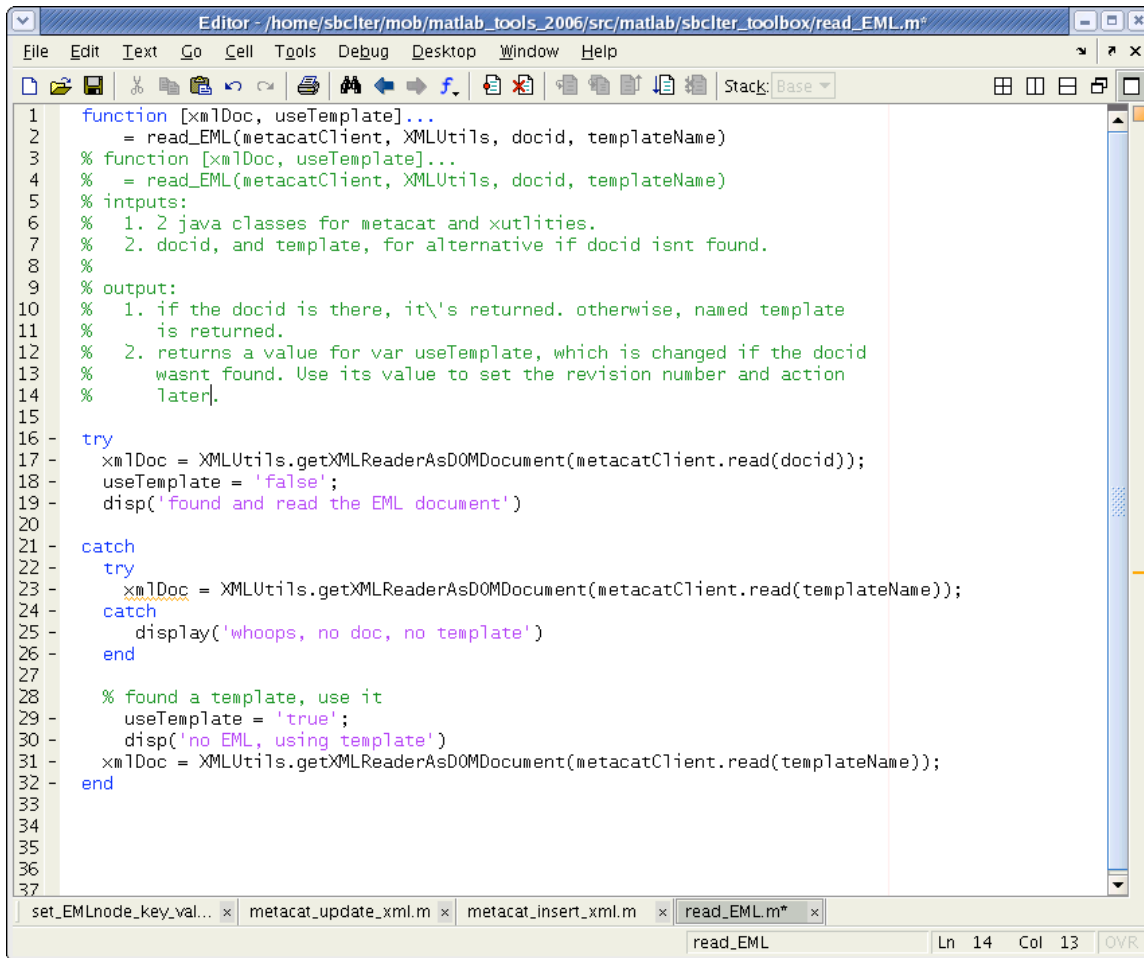
## 2.5 Data Export

SBC's data products are described in Ecological Metadata Language (Feagraus et al, 2005), and published on the SBC LTER website (<http://sbc.lternet.edu/data>), and in the network-wide LTER Data Catalog (<http://metacat.lternet.edu>). Researchers, technical staff and information managers outline the data products to be published. All data packages include table and attribute descriptions at a high level of completeness (LTER Network, 2004).

For the most part, the format of most of SBC's published data tables is stable. Our policy is to update tables with additional or replacement values, followed by update of the relevant metadata. Since metadata updates are often minor, these are accomplished in one of two ways. First, they can be handled manually by technical personnel. SBC uses the Oxygen XML Author<sup>®</sup> "tagless editor". We have created a customized framework "add-on" for EML documents and our metadata updates (SBC LTER, 2008). The Author<sup>®</sup> Editor isolates the user from XML markup, and the custom framework highlights EML metadata content expected to require editing. These features greatly ease use by those not trained in XML schema. Data packages updated this way are added to Metacat via a harvest list.

Alternatively, metadata updates and publication can be handled by external scripts. For example, another commonly used analysis language, Matlab<sup>®</sup>, has excellent XML integration. Matlab<sup>®</sup> also can import Java classes, allowing it to directly use the Metacat utility classes to login, read, and update EML instance documents in the catalog (Fig. 6). SBC also uses these functions (among others) to publish metadata for our other data products processed in Matlab<sup>®</sup>. We are working with researchers and information managers from a related project (the Partnership for Interdisciplinary Studies in Coastal Oceans, PISCO), to standardize and modularize this code for more data types (PISCO, 2008).

Ideally, we would like to export EML from our SAS<sup>®</sup> code as we do with Matlab<sup>®</sup>, and more completely integrate the final steps of data processing with data and metadata publication. However, the currently available version of SAS<sup>®</sup> is not able to access the XML Document Object Model (DOM) in a manner which allows either straightforward updates of EML instance documents, or the creation of new data packages.



```
1 function [xmlDoc, useTemplate]...
2 = read_EML(metacatClient, XMLUtils, docid, templateName)
3 % function [xmlDoc, useTemplate]...
4 % = read_EML(metacatClient, XMLUtils, docid, templateName)
5 % inputs:
6 % 1. 2 java classes for metacat and xutilities.
7 % 2. docid, and template, for alternative if docid isnt found.
8 %
9 % output:
10 % 1. if the docid is there, it\'s returned. otherwise, named template
11 % is returned.
12 % 2. returns a value for var useTemplate, which is changed if the docid
13 % wasnt found. Use its value to set the revision number and action
14 % later|.
15
16 - try
17 - xmlDoc = XMLUtils.getXMLReaderAsDOMDocument(metacatClient.read(docid));
18 - useTemplate = 'false';
19 - disp('found and read the EML document')
20
21 - catch
22 - try
23 - xmlDoc = XMLUtils.getXMLReaderAsDOMDocument(metacatClient.read(templateName));
24 - catch
25 - display('whoops, no doc, no template')
26 - end
27
28 % found a template, use it
29 useTemplate = 'true';
30 disp('no EML, using template')
31 xmlDoc = XMLUtils.getXMLReaderAsDOMDocument(metacatClient.read(templateName));
32 end
33
34
35
36
37
```

Figure 6: Screenshot of Matlab<sup>®</sup> code which uses Java classes to read an EML document so that specific parts of the DOM can be accessed. SAS<sup>®</sup> currently does not support such actions.

### 3. Results and Discussion

The SBC LTER maintains 10 to 15 datasets on kelp forest biota on a monthly to bi-monthly basis. Data processing standardization and automation are integral to keeping datasets current, clean and easily accessible to project investigators and the public. We have found a combination of Excel<sup>®</sup> spreadsheets and SAS<sup>®</sup> checking and processing loops to be optimal for this. Input forms are easily maintained by those responsible for (and therefore most familiar with) data collection. And because most undergraduate students are familiar with Excel<sup>®</sup> spreadsheets, their training time is reduced. Complex analysis code is best accomplished with a flexible scripting language such as SAS<sup>®</sup>, and ideally, the final data products can be integrated with scripts for publication in data catalogs.

One issue that we have not yet addressed adequately is that of flagging questionable or suppositious values where solid evidence of sampling error is lacking. The processing log that accompanies each data product is currently used to monitor suspicious observations and file maintenance events (Fig 5). SBC has collected standardized data for eight years, and we are developing a system of using accumulated knowledge to generate abundance and/or size thresholds for flagging suspicious values. Our community dynamics datasets encompass functionally and taxonomically diverse groups and the broad thresholds that we currently apply may require additional refinement. The use of scripting modules allows us to incorporate

relatively fine thresholds, even at the species level. We have not yet begun to investigate how this additional information could be incorporated into existing metadata, although a logical location for their descriptions is the attribute-level “methods” and “bounds” nodes.

Our current practices for updating metadata are not optimal because updates are somewhat disconnected from the data, and maintaining correspondence between the two requires human intervention or manual editing. Even though scripted updates can be automated, the production of reliable data packages still depends on data entities remaining consistently formatted and adhering to a predefined template. More complete integration between data table export and EML production is planned, but depends on a scripting language that adequately supports XML integration. And as our flagging methods and thresholds become more refined, access to the DOM from within SAS<sup>®</sup> will be not just desirable, but necessary.

Other planned improvements to coded export of EML metadata include replacing data package templates with entity-specific metadata that is imported into the processing script, and EML exported either to file system or inserted directly into Metacat. We are developing systems to store more metadata centrally, and build EML components rather than relying on static templates.

#### **4. Conclusion**

In long-term monitoring projects, well-defined data collection and processing protocols are highly recommended, if not required. Quality control in the field and during processing assures consistency and reliability of time-series data. The use of commercial desktop software for data entry and review has distinct advantages for an undergraduate work force. Because the needs of an evolving research project are subject to change, a scripting language such as SAS<sup>®</sup> is highly effective for processing, as its flexibility allows import and export from various sources. Other methods and interfaces certainly exist to regiment data flow, however the system described here is both efficient and suitable for managing SBC LTER’s time series data on kelp forest community dynamics.

#### **Acknowledgements**

US LTER Network, SBC LTER, Supported by National Science Foundation Award Nos. OCE-9982105, OCE-0620276

#### **References and software resources**

- Fegraus, E.H., S. Andelman, M.B. Jones, and M. Schildhauer. 2005. Maximizing the value of ecological data with structured metadata: An introduction to Ecological Metadata Language (EML) and principles for metadata creation. *Bull. Ecol. Soc. Amer.* 86:158-168. doi: 10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2
- LTER Network. 2004. EML Best Practices for LTER sites. <http://lternet.edu>
- O’Brien, M. C. and C. Burt. 2007. A Query Interface for EML dataTables. <http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/07spring/#4fa>
- PISCO. 2008. Software Development. <http://www.piscoweb.org/directory/development.jsp>
- SBC LTER, 2008. EML Framework for OxygenXML Author. [http://sbcdata.lternet.edu/external/SOFTWARE/OxygenXML\\_Author/](http://sbcdata.lternet.edu/external/SOFTWARE/OxygenXML_Author/)

## **MERCURY- A DISTRIBUTED METADATA MANAGEMENT, DATA DISCOVERY AND ACCESS SYSTEM**

**Giriprakash Palanisamy<sup>1</sup>, Bruce Wilson<sup>1</sup>, Ranjeet Devarakonda<sup>1</sup>, Jim Green<sup>2</sup>**

(1) Environmental Sciences Division, Oak Ridge National laboratory, Oak Ridge, TN

(2) On subcontract from Information International Associates, Oak Ridge, TN

### **Abstract**

Mercury is a federated metadata harvesting, search and retrieval tool based on both open source software and software developed at Oak Ridge National Laboratory. It was originally developed for NASA, and the Mercury development consortium now includes funding from NASA, USGS, and DOE. A major new version of Mercury was developed during 2007 and released in early 2008. This new version provides orders of magnitude improvements in search speed, support for additional metadata formats, integration with Google Maps for spatial queries, faceted type search, support for RSS delivery of search results, and ready customization to meet the needs of the multiple projects which use Mercury. For the end users, Mercury provides a single portal to very quickly search for data and information contained in disparate data management systems. It collects metadata and key data from contributing project servers distributed around the world and builds a centralized index. The Mercury search interfaces then allow the users to perform simple, fielded, spatial and temporal searches across these metadata sources. This centralized repository of metadata with distributed data sources provides extremely fast search results to the user, while allowing data providers to advertise the availability of their data and maintain complete control and ownership of that data.

**Keywords:** mercury, metadata management, data discovery, ornl daac, nbii

### **1. Introduction**

As the number of scientific datasets created by various research projects continues to increase, and both publishers and funding agencies require publication of the datasets as conditions for publications and funding, the number of places where users have to search to find relevant data sets also increases. Internet search engines (e.g. Google) are useful, but general purpose Internet search engines lack the specificity needed to search based on spatial, temporal, or other domain-specific keywords. Virtual observatories and distributed metadata search and data discovery systems are helping the scientists search those repositories to find and access the required data (Todd 2008). Distributed/virtual metadata systems typically harvest these metadata from various data providers and make it available through a single search system. In the mid-1990's, the NASA-funded Distributed Active Archive Center [REF: ORNL DAAC], at Oak Ridge National Laboratory (ORNL DAAC) developed a distributed metadata harvesting, search and data discovery system called Mercury [REF: Mercury], to search biogeochemical data archived at the ORNL DAAC. The Mercury system was developed to provide a single portal to information contained in disparate data management systems, and it has been improved and refined over time, with a major rewrite being completed during 2007. Mercury provides free text, fielded, spatial, temporal and keyword browse tree search capabilities. Mercury allows individuals and database managers to distribute their data while maintaining complete control and ownership. Mercury is also designed to be configurable to meet the needs of different



projects and different types of end users, and it is currently in use for a range of projects funded by the National Air and Space Administration (NASA), the US Geological Survey (USGS), and the Department of Energy (DOE). Mercury development is currently operated as a consortium, with the development and operating costs shared across these projects. In this paper we discuss Mercury's harvesting models, indexing techniques, and various search services that are available through the Mercury system.

## 2. Methods and Techniques

Mercury supports widely used metadata standards such as FGDC, Dublin-Core, Darwin-Core, EML and ISO-19115, and protocols and specifications such as XML and Z39.50. The new Mercury system is based on open source and Service Oriented Architecture and provides multiple search services.

The Mercury architecture includes different components, a harvester, an indexing tool, and a user interface. Mercury's harvester operates in two different models, 1) as a virtual internet database and 2) as a virtual aggregate database. The virtual internet database model organizes a new collection of data from informal systems spread across the internet. Typically, the data providers or the principal investigators create the metadata for their datasets and place these metadata in a publically accessible place such as a web directory or FTP directory. Mercury then harvests these metadata and builds a centralized index and makes it available for the Mercury search user interface.

In the virtual aggregate database model, Mercury harvests information from existing formal disparate database management systems (DBMS). In this model, where the metadata exists in remote databases, custom export programs can be easily written to extract the metadata from these DBMS. The metadata are then typically saved in xml files. Mercury then harvests the extracted metadata files and builds a centralized index for metadata searching (Figure 1). Some Mercury instances are using both these models to harvest the metadata. Mercury development team is currently working on enabling a metadata harvesting service using the Open Archives Initiative (OAI).

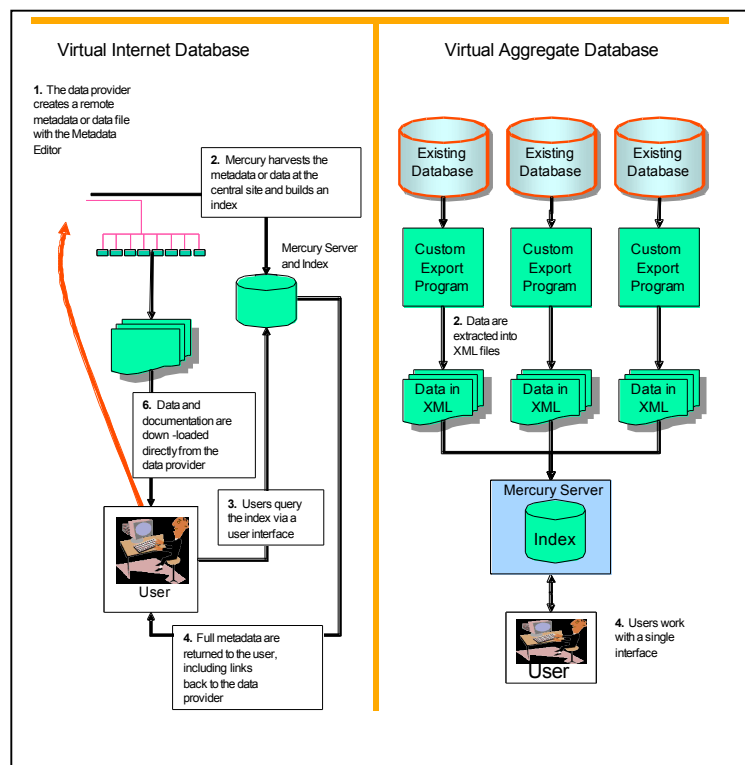


Figure 1. Mercury metadata harvesting

In the new version of Mercury, the indexing and searching interfaces were completely redesigned, eliminating the use of a commercial indexing tool, and replacing it with Lucene, which is an open source tool from the Apache project [REF: Lucene]. The new Mercury also

makes use of Solr [REF: Solr], which is an open source enterprise search server based on Lucene. Solr extends the functionality of Lucene, enabling much greater control over searching numeric types and dynamic fields, as well as enabling unique keys, and faceted searching. As an example, Solr gives the developer the ability to give special treatment to specific geotemporal coordinates. Special information that is used in an advanced search can be treated properly using Solr, as opposed to being buried among the competing rankings given by the Lucene to all the metadata content.

### 3. Results and Discussion

The typical Mercury user interface provides three different search capabilities. 1) simple search, 2) advanced search and 3) Web browse tree search. In the simple search option, users can perform a full text search. In the advanced search option, users will be able to search by specifying keywords, time period, spatial extend and the data provider information. Figure 2 is a snapshot of the Mercury advanced search interface used in ORNL DAAC [REF ORNL Mercury].

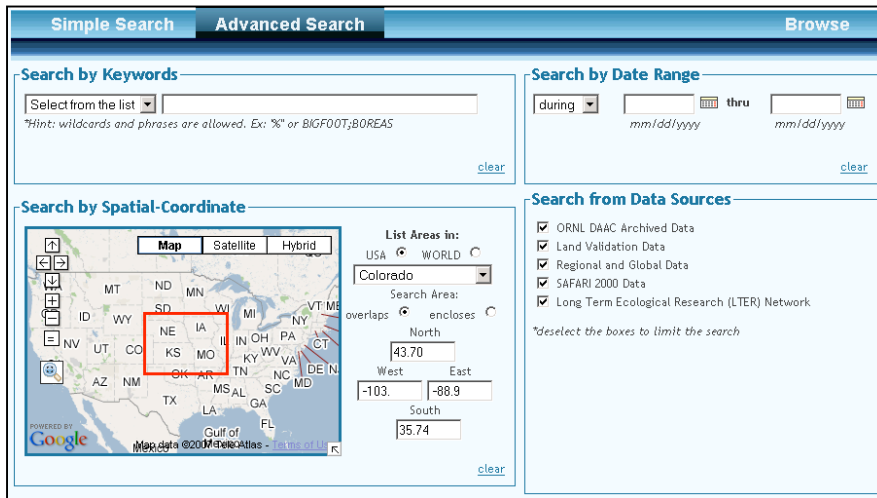


Figure 2. A snapshot of the ORNL-DAAC advance search interface

In the web browse tree search option, users will be able to drill down to their metadata of interest using a hierarchical keyword tree (figure 3).

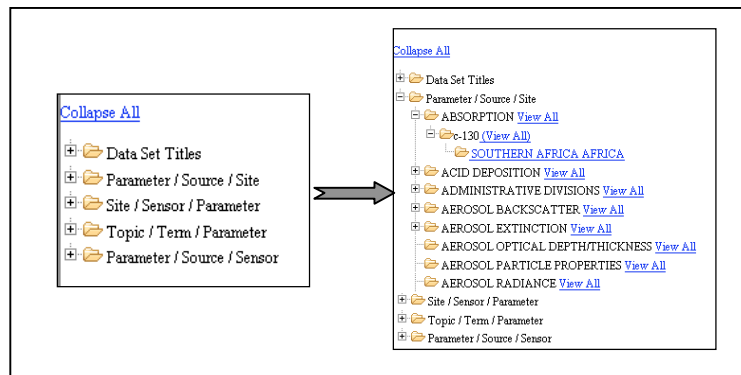


Figure 3. A snapshot of the ORNL-DAAC browse tree search architecture

Once the users enter their search criteria and perform the search, the results summary page displays the total number of records found for the search and option for filtering the search

results using logical groupings (by data providers, parameter, sensor, topic, project etc.). The summary page also allows the users to sort the results based on the search relevancy, period of record, source and project. The page shows push buttons in the top right to create an RSS feed, a bookmark or an email for these results. RSS or bookmarks enable refreshing the query matches periodically without the hassle of recreating the query.

The bottom of the summary page shows the results, snippets of the records that match the search/browse criteria, and a link to the full metadata and a link to access the associated data. The stars shown at the bottom of each record indicate the relative relevance of the matched criteria. The snippet includes the title and study date range, source provenance and excerpts from the abstract (Figure 4).



Figure 4. A typical look at the query results page

When the user clicks the “View Full Metadata” link found on the summary page, the Mercury metadata report’s page will be displayed. This page offers two styles to display a full metadata record. The Mercury by default offers a classic, well organized redux style at the full records page and additionally, it offers what it is known as the FGDC style, which would be very familiar to those who use the ESRI tools or that have used the previous mercury. It is plain text divided in 6 sections, with the underlying hierarchy preserved as indentation.

Users can create a bookmark, email their custom search results or subscribe to an RSS feed. RSS and bookmarks enable refreshing the query results periodically without the need to recreate the original query. For example, if the user searches for “soil temperature” in the LTER datasource, Mercury will return references to the 78 LTER metadata records which contain “soil temperature” in the metadata record. The user can then select the RSS button (Figure 5) on the

result summary page to get the RSS URL for this specific search criteria ( i.e., full text: soil temperature and data source: lter), an example url is something like:  
<http://mercdev3.ornl.gov/ornldaac3/send/processRss?term1=soil+temperature&term1attribute=txt&op1=&term6attribute=datasource&op6=+OR+&term6=lter&pageSize=10&start=0&sortattribute=default&sortattribute=default>. Users can then use this RSS URL in any of the many RSS readers (e.g., Google Reader, iGoogle, MyYahoo etc..) that are available online for subscribing to search results. Whenever the RSS reader refreshes the feed, Mercury will perform a new search and provide the latest search results, and the newly added records will be displayed at the top. The user can obtain the full metadata report by selecting the link found in the RSS feed.

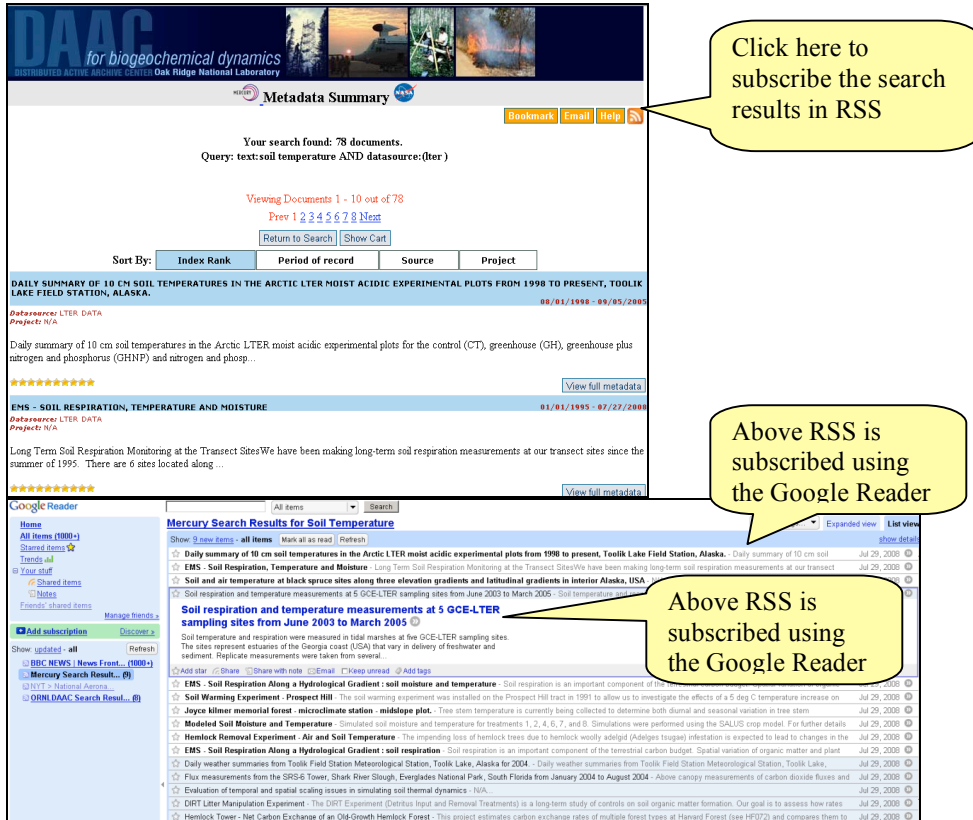


Figure 5. Creating RSS feed from a search result

Mercury also provides the harvested metadata to other applications (e.g., Google, NASA Global Change Master Directory, NBII Biobot). The National Biological Information Infrastructure [REF: NBII] Clearinghouse [REF: NBII CH] consumes the search results as portlets in their NBII portal web application, which is another way of displaying the customized search results in external web pages. Global Forestry Information Services [REF: GFIS] which is partnering with NBII Clearinghouse is harvesting all the forest related metadata records as RSS service and exposing those records through their search system.

#### 4. Conclusions

Mercury indexes and searches more than 50,000 metadata records through its various project-specific user interfaces. Mercury supports various metadata standards including XML, Z39.50, FGDC, Dublin-Core, Darwin-Core, EML, and ISO-19115. The new Mercury system is

based on open source and Service Oriented Architecture and provides multiple search services including; user interface search tools, RSS services for search results, bookmark search results, portlets supports.

### **Acknowledgements**

The Mercury consortium is funded by NASA, USGS, and DOE for a consortium of projects, including ORNL DAAC, NBII, DADDI, LBA, LTER, NARSTO, CDIAC, OCEAN, I3N, and IAI.

### **References:**

- Todd King, Narock, T, Walker, R., 2008. A brave new (virtual) world: distributed searches, relevance scoring and facets 1:29-34. doi: 10.1007/s12145-008-0002-7
- ORNL DAAC: <http://daac.ornl.gov/>
- Mercury: <http://mercury.ornl.gov>
- ORNL Mercury: <http://mercury.ornl.gov/ornldaac>
- Lucene: <http://lucene.apache.org/java/docs/index.html>
- Solr: <http://lucene.apache.org/solr/>
- NBII: <http://www.nbio.gov/>
- NBII CH: <http://mercury.ornl.gov/nbio>
- GFIS: <http://www.gfis.net/>

## LIVE FROM THE FIELD: MANAGING LIVE-IMAGE DATABASES AT THE VIRGINIA COAST RESERVE

John Porter and David E. Smith

Dept. of Environmental Sciences, University of Virginia

### Abstract

Imagery of ecological systems can be used to observe organisms, to observe rare events and to document ecological changes in the system. Here we describe the Virginia Ecomcam System which uses wireless networks and web cameras to capture imagery from remote barrier islands. The system uses a MySQL database, PHP code and shell scripts to generate custom displays, including animation, browse, pan and change detection displays, for analysis.

**Keywords:** image database, webcam, wireless network, barrier island

### 1. Introduction

The dictum “one picture is worth ten thousand words” (Bernard 1927) is especially true when monitoring ecological systems that incorporate diverse drivers, some of which may have been unanticipated when the monitoring program was initiated. The Virginia Coast Reserve Long-Term Ecological Research (VCR/LTER) project (<http://vcr.lternet.edu>) has as its focus the ecology of the relatively pristine barrier islands off the coast of the Delmarva Peninsula (Figure 1). The barrier island/lagoon system is a logistically challenging environment for both researchers and instruments, characterized by salt water and salt spray, large but shallow bays, and blood-sucking insects (Lee 1832, Hayden et al. 1991, Erickson and Young 1995, Oertel and Overman 2004).

During times of extreme weather events, such as hurricanes or during the winter when the bays can freeze over, it is not possible for researchers to physically visit their research sites. For these reasons, we sought to develop a wireless network incorporating network cameras, using commercial-off-the-shelf components, that link the barrier islands to one another and to our laboratory on the mainland (Porter 2007), and to provide the resulting images to researchers. Wireless networks allow the collection of data over broad spatial extents at high frequencies permitting unobtrusive observation and remote

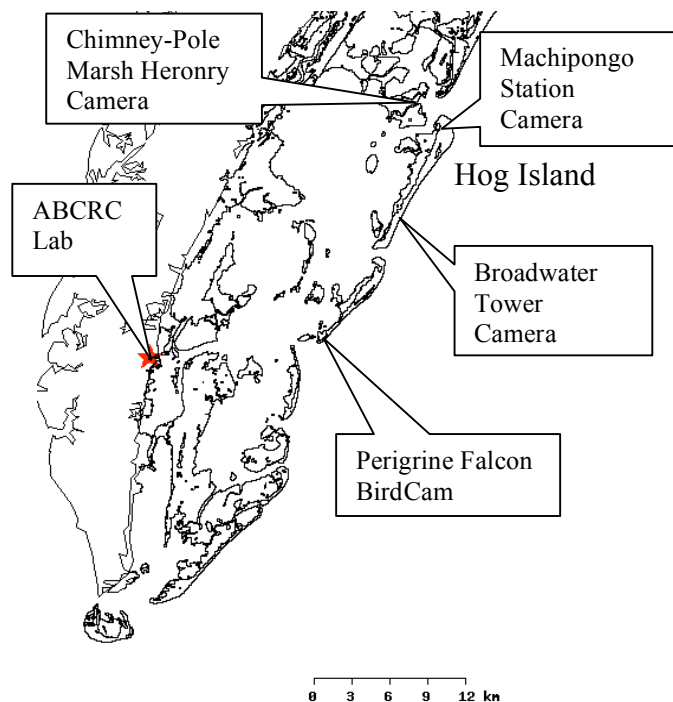


Figure 1: Web Cameras on the Virginia Coast Reserve.

control of robotic sensors (Arzberger et al. 2005, Porter et al. 2005, Hart and Martinez 2006).

Our primary objective in creating the Virginia Ecocam System was make information contained in images accessible to ecological researchers, most of whom have little background in digital image processing. To accomplish this objective, we concluded that our system should:

1. automatically capture and ingest of images from multiple web cameras,
2. query images based on location and time,
3. provide tools for image browsing, animation and display, and
4. provide change analysis tools for selected imagery.

## **2. Methods and Techniques:**

The Virginia Ecocam System is comprised of three major components: a wireless network, web cameras and a data system. During 2002, we developed a network backbone that connects the Anheuser-Busch Coastal Research Center (ABCRC) laboratory on the mainland to Hog Island, 22 km distant, via a proprietary 900 MHz network radio at 3 Mbs. On the islands, access is provided by amplified Wi-Fi (802.11b,g) access points at two major nodes located at Broadwater Tower and Machipongo Station at both ends of Hog Island, which also host major camera installations (Figure 1). The nodes on Hog Island provide broad coverage of the adjoining barrier islands and bays. The resulting network has been used to support a wide variety of activities including: real-time observations of hurricane-driven flooding, monitoring bird foraging, collecting meteorological and tide data, observing nesting peregrine falcons, calibrating and configuring a tunable diode laser trace gas analyzer, radar-tracking of migrating birds, and even videoteleconferencing (Porter 2007).

Cameras vary in mission, location and capabilities and frequency of data collection. Table 1 contains a list of major image sources, the dates and frequency of image collection and the number of images archived. Two camera installations, located in towers, focus on large-scale landscape change including tidal and storm-driven flooding, plant phenology and human impacts. Other cameras are focused on specific organisms of interest. Submersible “CrabCams” monitor fiddler crab habitat and populations, while a pan-tilt-zoom camera provides mosaics of images for monitoring habitat use by wading birds. A series of FalconCams operated by the Center for Conservation Ecology at the College of William & Mary and their collaborators monitor the nests of peregrine falcons.

Depending on its purpose, each camera is used to provide a high-frequency feed for live observation, with periodic archiving of images for retrospective analyses. Live feeds provide updated images every 2-10 seconds which is a compromise between researcher needs, camera capabilities and network bandwidth. These high-speed feeds are used by bird researchers to check on the current condition of chicks or to monitor habitat use. Archive images are typically taken hourly, although during some periods of higher interest, they may be taken with greater frequency.

Many network cameras (“webcams”) support two modes for image capture. In a “push” mode they will use the File Transfer Protocol (FTP) or electronic mail to send images at set times, or when events occur. In a “pull” mode, a web browser can request an image and download it. For our collection efforts, we depend almost entirely on “pull” connections. This is

Table 1: Web cameras harvested by the Virginia Ecocam System. All cameras have a resolution of 704x480, with the exception of the Chimney-Pole Marsh camera which has a resolution of 1284x950.

<i>Camera Locations</i>	<i>Cameras</i>	<i># Images per Harvest</i>	<i>Frequency of “Live” Data/ Bandwidth Required</i>	<i>Frequency of Archival Collections</i>	<i>Archived Dates</i>	<i>Number of Images Archived</i>
<b>Broadwater Tower – Landscape</b>	1 PTZ	20	5 seconds / 3 KB/sec	Hourly	April 2002-present	593,157
<b>Machipongo Station – Landscape and CrabCams</b>	1 PTZ, 1-3 fixed	6-8	10 seconds / 6 KB/sec	Hourly	April 2003-present	88,339
<b>Machipongo Station Landscape/Birds</b>	Scan of PTZ camera	132	N/A	Bi-Hourly	July 2006-present	558,134
<b>Chimney-Pole Marsh – Heronry</b>	1 fixed	1	5 Hz / 1 MB/sec	Variable (change detection)	July 2003-August 2003	10,329
<b>Cobb Island - FalconCam</b>	1 PTZ, 3 fixed	2-4	2-10 seconds / 12 KB/sec	10-minutes to Hourly	May 2005-present	70,819 (stored but not archived)
<b>Other FalconCams</b>	3 PTZ, 9 fixed	1-4	5-10 seconds / 36 KB/sec	Not archived	Not archived	Not archived

primarily due to our use of pan-tilt-zoom (PTZ) cameras. These cameras can be controlled using configuration strings sent via a web browser. However, they typically do not have a mechanism for reporting their current direction, tilt or zoom level, so control of the camera must be closely coupled with capture of images if images from known locations are to be obtained. Unix shell-scripts controlling the capture operation are run at pre-scheduled times using the standard Unix CRONTAB scheduler. A shell script will point a camera at a location, order it to focus, then capture an image, before repeating the process for a sequence of locations. Additional scripts produce short-term products (e.g., animations of the last three days) and place images in a temporary, but web accessible, holding area. Periodically a PERL script is used to move images



out of the temporary holding area to our permanent archive. High-frequency live images are collected using set of continuously-running shell scripts that use a command-line web browser to save images to a file, delay for a set period and then repeat.

The permanent archive consists of a MySQL database (<http://mysql.org>) with tables for cameras, positions, images, and annotations (Figure 2). The “camera” table contains

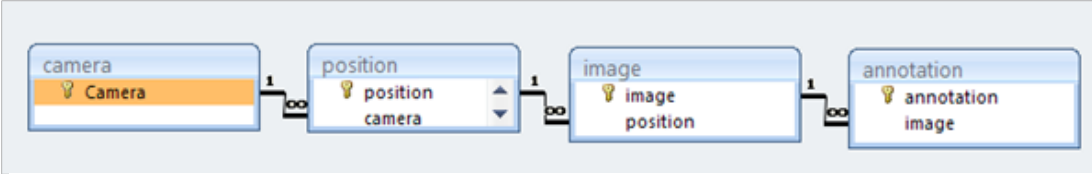


Figure 2: Schematic Entity-Relationship Diagram for the MySQL database. The “image” table has the highest volume as most images are not annotated.

information on the name, network address and type of camera. The “position” table contains information on the names and locations of particular “views” provided by each camera. Fixed cameras have a single position, but pan-tilt-zoom cameras may have up to 20 positions. The “images” table contains the details on individual images including date, size and format, along with the location where the actual image file is stored in the file system. The image files themselves are not stored directly in the database because as highly compressed, binary objects they are easily retrieved from a time-structured (position, year, month) file system, and including them directly into the database would greatly increase the size of database backups from a few hundred megabytes to many gigabytes. The file system for images is maintained on a network appliance using RAID 5 with periodic off-site backups. Finally, the “annotations” table contains additional information from users about individual images, such as comments on image content or quality.

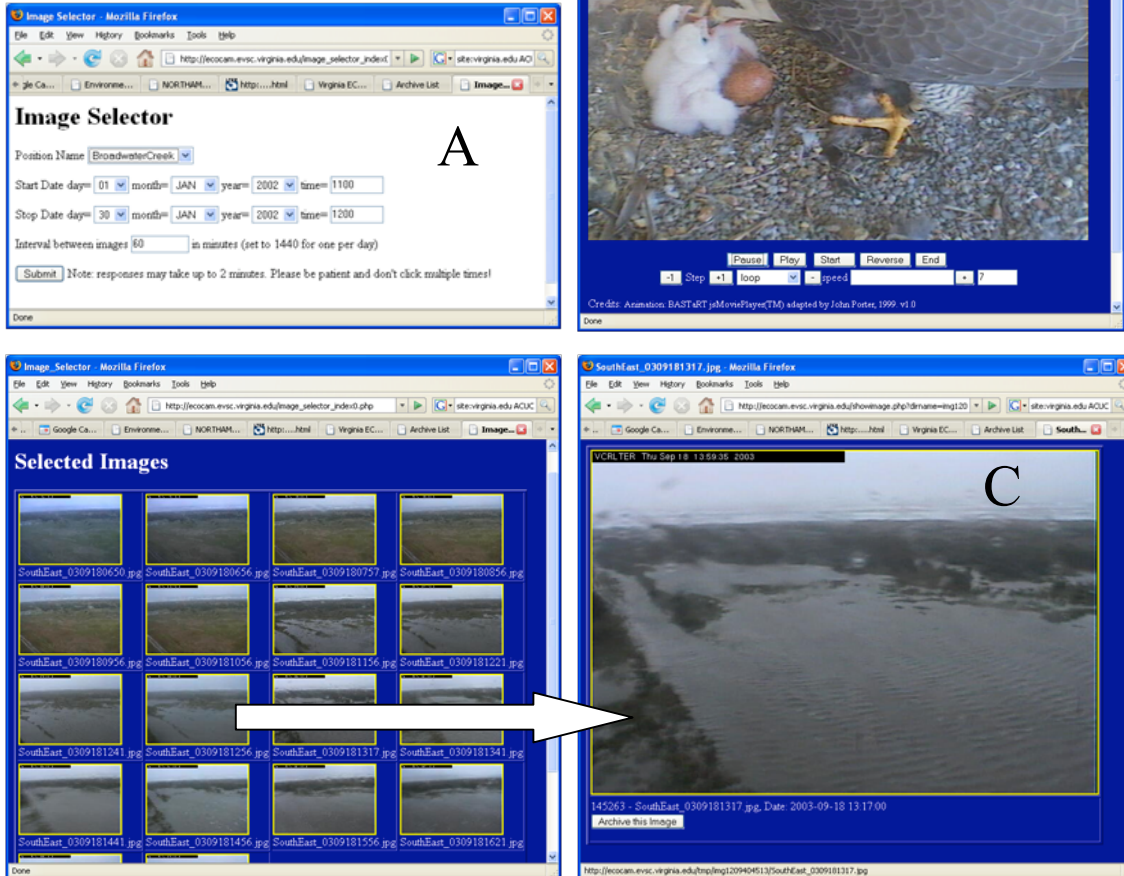
### 3. Results and Discussion

To facilitate use of the images by ecological researchers, we have provided a number of query-based tools that allow a researcher to efficiently select and view images based on location and time (Figure 3). A variety of tools are required to support different modes of inquiry. In some cases researchers are interested in obtaining representative images of sites, in others the changes between images are of more interest, such as a sequence of images showing the deposition (and subsequent removal) of wrack on a marsh surface.

Web forms (Figure 3A) coupled to the MySQL database using PHP programs allows users to specify the location, camera position, start and stop dates, times of day and the interval between successive images. Images are then sequentially retrieved based on best approximation of the desired time(s). Sequential retrieval is required because exact matches (accurate to the second) are unlikely.

Once retrieved, the images selected are then displayed either as an animation, using a JavaScript image viewer which includes controls for adjusting playback speed, or single-stepping backwards and forwards through images (Figure 3B). This allows users to visually detect changes and to isolate those changes down to individual frames. Alternatively, images can be displayed simultaneously using an “index frame” page (Figure 3C). This allows a user to rapidly browse for images containing features of interest and to obtain full-resolution images.

Figure 3: Generic user interfaces: A- a search form, B- An animation display, and C- An index page that allows a user to expand selected images of flooding during a hurricane by clicking on them.



Some features of interest may be hard to spot. For example, one use of network cameras has been the monitoring of populations of fiddler crabs (*Uca pugnax*), which are cryptic and difficult to identify from still images. We therefore built a simple interface that employs easy-to-use image processing tools to detect areas of change between temporally adjacent images (based on a user-selectable threshold and a simple differencing algorithm) to help users identify the locations of crabs (Porter 2005). Users select images to compare from an index-page display and can adjust the degree of discrimination. They then receive a processed image where areas of change are highlighted (Figure 4).

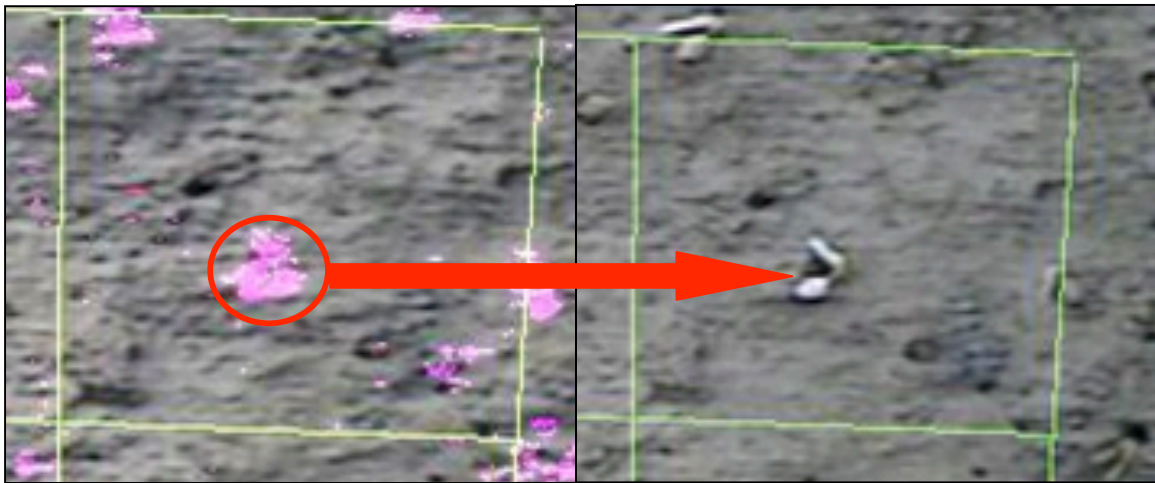


Figure 4: Highlighted change in processed image (circle, left) identifies location of an otherwise cryptic fiddler crab (right). A superimposed 25x25 cm reference frame helps users to enumerate fiddler crabs.

In other cases, features of interest are highly visible, but are broadly distributed across the landscape. For example, egrets are large white birds that show up well, even at a distance. However, they forage broadly in the barrier island wetlands, so that any particular camera view is not likely to have a bird in it. We therefore designed a tool that provides a mosaic, low-resolution view of all the areas visible from the Machipongo Station camera. This allows a quick scan by a user to detect if birds are present. Individual frames can then be blown up to full resolution to provide details of the birds present (Figure 5).

#### 4. Discussion

There are a wide array of software products and systems focused on managing large image repositories and providing them via the World-Wide Web, including MorphBank, Gallery and FEDORA. MorphBank (<http://morphbank.net>) is representative of projects that focus on particular types of images, specifically images of biological specimens and organisms. It includes sophisticated search tools that take advantage of the extensive metadata provided with each image (taxon, specimen, view, collection, publication and locality).

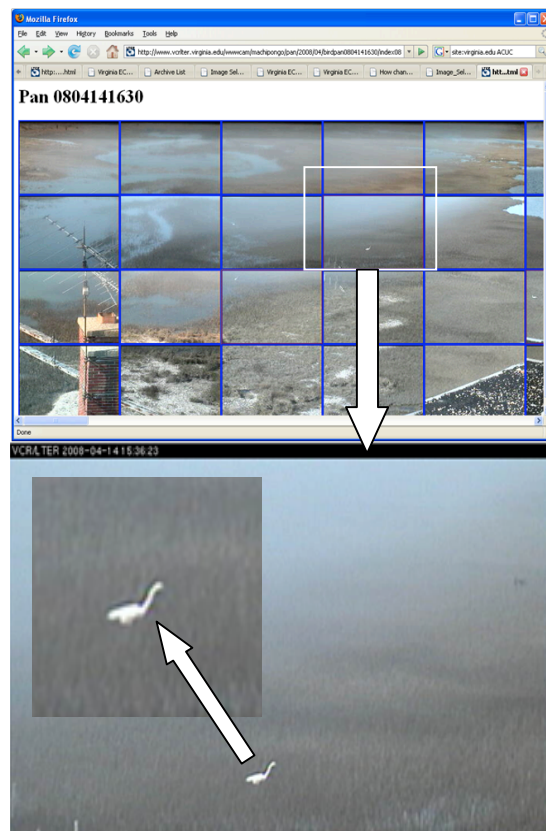


Figure 5: A mosaic of thumbnail images for browsing (top) to locate high-resolution images for detailed inspection (bottom and inset).

Gallery (<http://gallery.menalto.com/>) is typical of general tools for organizing images, with a focus on manual uploads. In contrast to MorphBank (and even the Virginia Ecocam) it has virtually no mandatory metadata, although it supports titles, captions, locations, times and keywords for individual images and collections of images. The Flexible and Extensible Digital Object and Repository Architecture (FEDORA) (Payette and Lagoze 1998) is typical of more generalized digital library systems that can be adapted for image data (Johnston 2005). Like the more specialized MorphBank, these systems are heavily metadata dependent, but can be customized to meet specific purposes.

The Virginia Ecocam System draws on elements of all these systems. The index pages are similar to those found in Gallery, query capabilities are similar to MorphBank, whereas the annotation functions are similar to FEDORA. It also draws on elements found in weather web sites (animation viewers) and astronomical systems (change detection).

As with any system, there are elements that could be improved. One of these is coupling the change detection, used to look at fiddler crabs (Figure 4), to the large pans aimed at detecting egrets (Figure 5). Change detection could then facilitate semi-automated detection of egrets. Another is improving the scalability of the system. Asynchronously running harvest scripts could lead to unexpected bottlenecks in resource use if a large number of additional cameras were added (with the current number of cameras, the load is so low that bottlenecks are not a significant problem). However, developing an image harvest system that uses a single controller is complicated because of the latency involved in some of the operations, such as panning or focusing cameras, coupled with the high rate of data acquisition.

## 5. Conclusions

Given the wide availability of systems designed for managing images, there seems little justification for creating yet another one. However, the Virginia Ecocam System is needed because of the unique mix of capabilities needed to support ecological research. Moreover web cameras, because they are controllable, present opportunities for combining images (both temporally and spatially) that are not available from more eclectic image sources.

## Acknowledgements

This material is based upon work supported by the National Science Foundation under Grants DEB-0080381 and DEB-0621014 and by a grant from the Virginia Environmental Endowment. The Virginia Coast Reserve of the Nature Conservancy provided access to study sites. Thomas Williams and David Hughes provided aid in establishing the wireless network and in the installation of web cameras.

## References

- Arzberger, P., J. Bonner, D. Fries, and A. Sanderson. 2005. Sensors for Environmental Observatories: Report of the NSF Sponsored Workshop December 2004. World Technology Evaluation Center (WTEC), Inc., Baltimore, MD.  
[http://wtec.org/seo/final/Sensors\\_for\\_Environmental\\_Observatories.pdf](http://wtec.org/seo/final/Sensors_for_Environmental_Observatories.pdf)
- Bernard, F. R. 1927. "Make a cake for Bobby". Pages 96-97 in *Printers' Ink*.  
<http://www2.cs.uregina.ca/~hepting/research/web/words/27.gif>
- Erickson, D., and D. R. Young. 1995. Salinity response, distribution, and possible dispersal of a barrier-island strand glycophyte, *Strophostyles Umbellata* (Fabaceae). *Bulletin of the Torrey Botanical Club* **122**:95-100. <Go to ISI>://A1995RK72900001

- Hart, J. K., and K. Martinez. 2006. Environmental sensor networks: a revolution in earth system science? *Earth Science Reviews* **78**:177-191.
- Hayden, B. P., R. D. Dueser, J. T. Callahan, and H. H. Shugart. 1991. Long-term Research at the Virginia Coast Reserve: Modeling a Highly Dynamic Environment. *Bioscience* **41**:310-318.
- Johnston, L. 2005. Development and Assessment of a Public Discovery and Delivery Interface for a Fedora Repository. *D-Lib Magazine* **11**.
- Lee, R. E. 1832. Letter from Smith Island. *in* B. M. Barnes and B. R. Truitt, editors. *Seashore Chronicles: Three Centuries of the Virginia Barrier Islands*. 1997. University of Virginia Press, Charlottesville, VA.
- Oertel, G. F., and K. M. Overman. 2004. Sequence morphodynamics at an emergent barrier island, middle Atlantic coast of North America. *Geomorphology* **58**:67-83. <Go to ISI>://000220353100004
- Payette, S., and C. Lagoze. 1998. Flexible and Extensible Digital Object and Repository Architecture (FEDORA). *in* *Research and Advanced Technology for Digital Libraries*. Springer, Berlin / Heidelberg.
- Porter, J. H. 2005. Life at the command line. *LER Databits* **Fall 2005**.  
<http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/05fall>
- Porter, J. H. 2007. From the Barrier Island to the Database: Evolution of wireless sensor networks on the Virginia Coast Reserve. *in* *Conference on Coastal Environmental Sensor Networks*. Center for Coastal Environmental Sensor Networks, University of Massachusetts, Boston, Boston, MA, USA.  
<http://cheetah.cs.umb.edu/ocs/viewabstract.php?id=33>
- Porter, J. H., P. Arzberger, H.-W. Braun, P. Bryant, S. Gage, T. Hansen, P. Hanson, F. P. Lin, C. C. Lin, T. Kratz, W. Michener, S. Shapiro, and T. Williams. 2005. Wireless Sensor Networks for Ecology. *Bioscience* **55**:561-572.

## **AN OVERVIEW OF QUALITY CONTROL PROCEDURES FOR BUOY DATA AT THE NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION'S (NOAA) NATIONAL DATA BUOY CENTER (NDBC).**

**Ian Sears**

National Data Buoy Center, Stennis Space Center, MS

### **Abstract**

The National Data Buoy Center collects, quality controls, and disseminates, in near real time, approximately six-million observations per year from meteorological, oceanographic, and water level sensors from over 500 moored buoys and coastal stations. This paper explores the National Data Buoy Center's quality control process and planned improvements.

### **1. Introduction**

NDBC applies quality control algorithms to data in near real time, and manually examines the data using computer generated graphics and tools within twenty-four hours of acquisition, as well as a final monthly review before data are archived at the National Climatic Data Center (NCDC) and the National Oceanographic Data Center (NODC).

NDBC operates more than 105 moored buoys and 55 Coastal-Marine Automated Network (C-MAN) stations. This network consists primarily of moored buoys and shore and platform-based coastal marine stations around the continental U.S., Alaska, Hawaii, and the Great Lakes. C-MAN stations are located at coastal locations such as lighthouses, piers, and offshore navigation platforms. These stations collect the same meteorological parameters as buoys, and oceanographic parameters as proximity to water will allow (Conlee and Moersdorf, 2004). NDBC also ingests, quality controls, and disseminates data from Integrated Ocean Observing System (IOOS) and National Ocean Service (NOS) stations. IOOS and NOS stations include moored buoys and coastal stations from many different types of organizations such as academia, other NOAA and federal activities, research organizations, and private industry. Given that meteorological forecasts and analyses, warnings, model initializations, research, and important decision making rely on data from buoys and coastal stations, it is imperative that NDBC disseminate in near real time, and archive accurate data.

### **2. NDBC system and sensor background**

NDBC uses different buoy technologies to collect data from a wide range of coastal environments. The NDBC-developed 3 meter (m) discus buoy is a staple of most coastal locations. NDBC also operates similar 2.4m discus buoys. The 6m NOMAD buoy of Navy 1950's heritage is used farther offshore and in harsher environments. The large 10 and 12m buoys that must be towed to station locations are typically used in extremely remote and harsh areas where servicing intervals may be up to 4 years. Buoys located in remote and harsh environments contain complete redundant instrumentation for survivability (Conlee and Moersdorf, 2004). Sensors on 6m NOMAD buoys and C-MAN stations cannot take directional wave measurements. Because of this limitation, NDBC developed 1.8m and 1.5m foam hull Coastal Oceanographic Line of Sight (COLOS) buoys to be co-located with select stations that are unable to collect directional wave data (Conlee and Moersdorf, 2006). NDBC recently made this buoy technology operational.

Most 3m and all 6m buoys have aluminum hulls. NDBC developed value-engineered buoys which have foam hulls. These hulls are on 2.4m and 3m discus buoys. 10 and 12m buoys have steel hulls. The survivability of the foam and aluminum hulls is less than the sturdier steel hulls, however, they are much more cost-effective. More information on the NDBC Moored Buoy Program can be found at <http://www.ndbc.noaa.gov/mooredbuoy.shtml>. Figures 1 and 2 show typical stations.

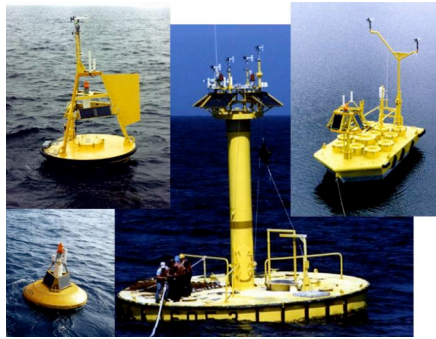


Figure 1: Clockwise from top left. 3m and 10m discus, 6m NOMAD, 1.8m COLOS



Figure 2: Typical C-MAN stations

DB  
C  
ope

rated stations and some IOOS stations employ duplicate sensors for air temperature, humidity, barometric pressure, and winds. The duplicate sensors act as a secondary sensor for the station in the event that the primary sensor fails, and as an accuracy comparison sensor. Analysts evaluate daily, and release data from the duplicate sensor that is determined to be more accurate. NDBC operated stations, which are difficult to service due to location or environmental conditions, use duplicated systems on board. These buoys have two independent, on board computer systems known as payloads, each with their own suite of sensors. The payload creates a raw message with all measurements taken by the station. Data are transmitted from buoys and C-MAN stations via satellites. NDBC uses Geostationary Operational Environmental Satellite (GOES), Iridium, and Service ARGOS to transmit data. Iridium and Service Argos provide communications services to the public, where as GOES is primarily reserved for government activities. The data are routed to the NWSTG where automated Quality Control (QC) algorithms are applied. The QC algorithms are written in C, and are designed to work within NDBC's database structure.

NDBC and IOOS partners have developed the ability to make a variety of measurements including atmospheric pressure, wind speed and direction, air and water temperature, wave energy spectra (directional and non-directional), relative humidity, ocean current velocity, precipitation, salinity, solar radiation, visibility, water level, and water quality (NDBC, 2003).

### 3. Automated Near Real Time Quality Control

All data from sources previously mentioned arrive, by different means, on two redundant UNIX servers running simultaneously at the National Weather Service (NWS) Telecommunications Gateway (NWSTG). Software on these servers start the QC process and data processing in near real time by applying algorithms, inspecting transmission quality, and giving instructions on the release of data. The redundant servers provide automatic backup

capability without the need for manual intervention if a workstation fails or the processing crashes (Gilhousen, 1998).

NDBC uses both simple and complicated QC algorithms. Simple algorithms include range, rate of change, and consistency checks. More complicated algorithms include comparison of related parameters such as wind speed and wave height, and the detection of high frequency spikes in spectral wave data (NDBC, 2003).

The software applies either hard flags or soft flags to the data that fail automated QC checks. Hard flags prevent data release while soft flags instruct human data quality analysts to investigate the data further. Capital letters and lower-case letters represent hard and soft flags respectively to analysts. Figure 3 illustrates an instance where WSPD1 and WSPD2 (wind speed sensor 1 and wind speed sensor 2) were automatically soft flagged with a k flag, because the difference between the two sensors was too large. The analyst manually instructed the NDBC computers to no longer release WSPD1 with a D flag and release WSPD2 instead. A description of all flags can be found in the *Handbook of the Automated Quality Control Checks and Procedures of the National Data Buoy Center* (NDBC, 2003.) This figure also illustrates pre-generated plots described later

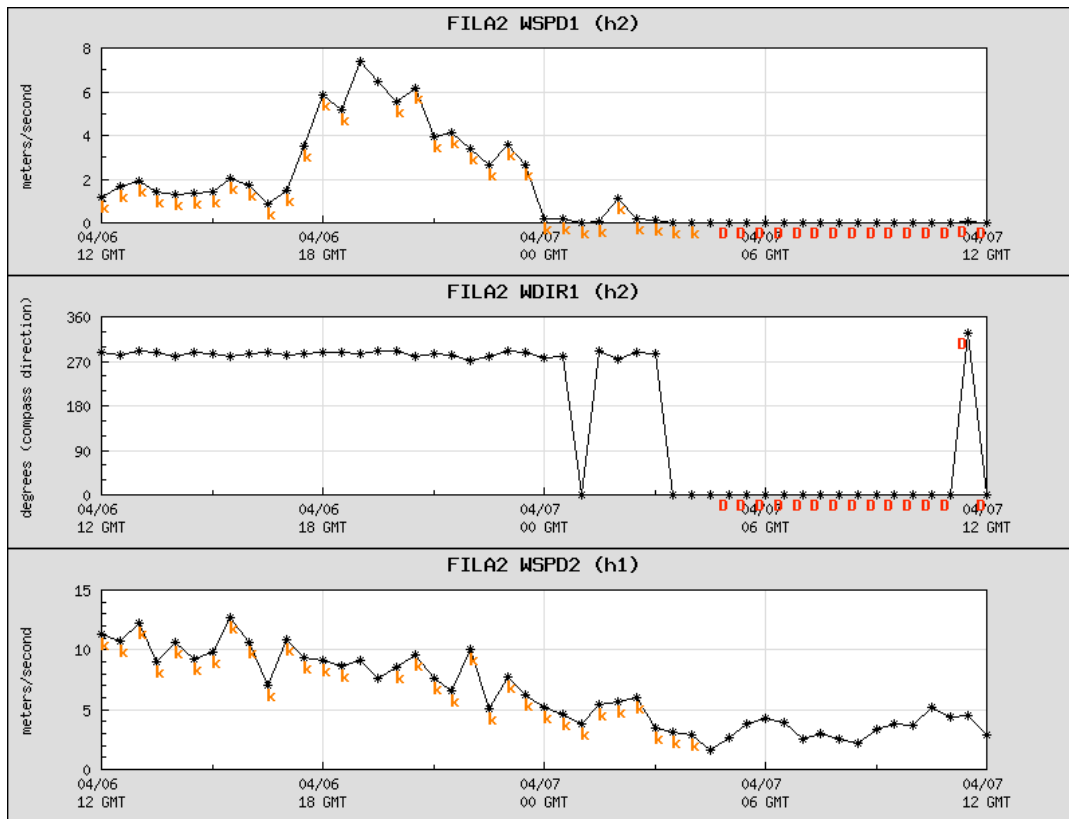


Figure 3: Typical pre-generated time series plot with both soft and hard flags.

Analysts update a control file that stops the release of data from degraded sensors after they are detected by manual analysis or the automated QC process. The control file then applies hard flags to all subsequent data during the real-time processing at NWSTG (Gilhousen, 2003).

To ensure the release of crucial data, during unusual weather situations, such as hurricanes and severe winter storms, analysts instruct the servers to release data, in near real



time, that would fail automated QC checks such as range and rate of change limits (NDBC, 2003).

NDBC inspects transmission quality and errors such as truncations to the message or parity errors. The NDBC computers will try to decode data that have transmission errors. Algorithms flag data as “missing” to measurements that cannot be decoded as a result of transmission errors (NDBC, 2003).

#### **4. Manual Near Real Time Quality Control**

After applying all QC algorithms and instructions, NWSTG routes all data, including hard and soft flagged data, to NDBC at Stennis Space Center, Mississippi, but only routes data that have passed all automated QC checks, without any automated QC flags, to the end user via the Global Telecommunications System (GTS), NOAAport, Family of Services, and eventually the NDBC website. The data received at NDBC populate an Oracle database allowing analysts to view data in its entirety, including the raw message<sup>11</sup> transmitted from the station, and all processed data complete with flags. Analysts at NDBC review data transmitted from all stations within twenty-four hours of acquisition. Analysts use computer-generated graphical displays that show all flags generated by the automated QC process. The graphical displays help analysts identify the often subtle degradation in systems and sensors (NDBC, 2003). Analysts view processed data through commercial off-the-shelf weather analysis software. Data received via a NOAAport satellite are displayed geographically in standard surface weather report symbols and analyzed in relation to satellite and numerical model data.

The present manual QC process consists of many different independent and related activities. Analysts print and inspect daily, all of the processed data for a twenty-four hour period in tabular form with any flags printed next to the failed or suspect data. Analysts look for gross errors missed by the automated QC process, erroneously flagged data, soft flags for further inspection, and subtle degradations. Analysts use pre-generated time-series plots to inspect the data from the previous twenty-four (Figure 3) and seventy-two hours. Data on the time-series plots include wind speed, gust, and direction, barometric pressure, air and water temperature, dew point, relative humidity, wave height, period, and direction, solar radiation, and system health parameters such as battery voltage and current. In addition to viewing sensor data independently with flags, analysts can define related measurements to be plotted on the same time-series plot. For instance, air temperature can be plotted against the air temperature of a nearby station or against numerical model data. Analysts may choose to view air temperature and wind direction on the same graph. A sharp change in wind direction and air temperature at the same time may correlate to a frontal passage, thus giving physical justification for a sharp rate of change of a parameter. The analysts have the flexibility to relate any set of available parameters to each other allowing for an efficient evaluation of data.

NDBC manually validates ocean data. Analysts currently use the NDBC web site to view oceanographic time-series plots. Viewing data this way does not allow the analysts to view the automated QC flags. The analysts must query the Oracle database to view flags generated by automated QC.

In addition to pre-generated plots, analysts have the ability to view custom time series, buoy position, stick, scatter, spectral wave density and direction, and wave energy vs. wind speed plots. Custom plots provide analysts a more dynamic and in depth look at data when pre-generated aids do not provide enough information about the health of a system or sensor.

---

<sup>11</sup> Raw data is only received from NDBC operated stations.

Analysts continuously display data using commercial weather analysis software. The displays allow, at a glance, a quick and gross check of certain parameters transmitted by stations. Analyst display pre-generated regional displays that run on a 3 hour loop and automatically update. A continental 12 hour loop, which encompasses the majority of the buoy network, exists as well. Once every two hours, analysts review all the regional loops looking for sensor, system, or communication failures. The purpose is to find problems sooner than the daily evaluation could detect them. For instance, if a station displays a wind from the north, and all surrounding stations display winds from the south, analysts will quickly notice, investigate, and take action if required on the discrepancy.

Analysts monitor communications hourly by viewing a report that indicates the length of time since the last data were received at NDBC. A daily report is also generated that indicates, for each station, which hours of the day data were not received. Analysts use the looping displays to help recognize communication issues. The dot on the display representing a buoy changes color from blue to bright orange when data have not been received in over two hours. Communications issues can occur on the buoy, hardware that receive the data, or anywhere in between, i.e. satellites, phone lines, NWSSTG server outages, etc. Once a communication issue has been identified, analysts will take the proper course of action depending on where the suspected communication failure occurs. The analysts will notify the Information Technology department if the suspected communication issue is shore-side, and write a Station Discrepancy Report (SDR) if the communication problem lies on the buoy.

When a sensor or system fails, analysts write SDRs which give vital information to the operations department and engineers. This information notifies the operations department that a system or sensor has a discrepancy and that action is required. SDRs provide time frames and history of system or sensor failures. The information is used to assist in the design of more reliable sensors and systems.

## **5. Pre Archive Quality Control**

Prior to archival, all data from NDBC operated stations go through a final check. This process is undertaken to find erroneous data that have been overlooked due to the limitations of the daily QC process. Figure 4 shows an air temperature data point that was clearly degraded. Viewing a months worth of data at once easily allows the analyst to identify and remove from the archive the data point. Only data that have met NDBC standards for accuracy, are archived (NDBC, 2003). NDBC stated accuracies can be found at <http://www.ndbc.noaa.gov/rqa.shtml>.

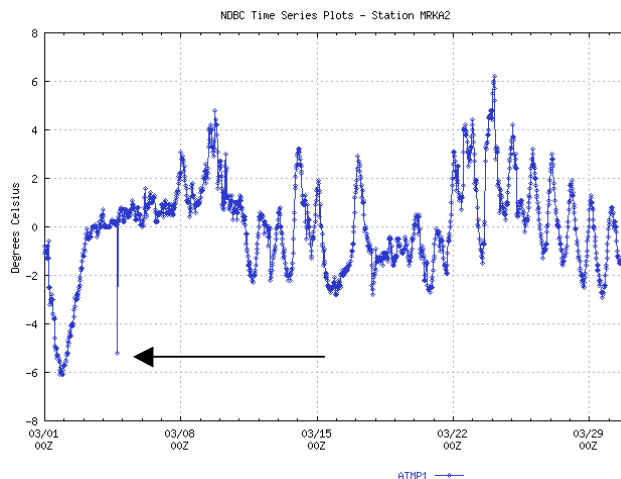


Figure 4 The arrow points to an air temperature data point that was found to be in error during the monthly QC check prior to archival.

## 6. Planned improvements to the NDBC Quality Control process

NDBC continues to work on improving the QC process. NDBC technical support contractors are working to streamline some procedures previously mentioned, such as creating pre-generated time-series plots for oceanographic data, and eliminating paper printouts of data by providing the same data on a computer screen complete with color coded flags. Scientists continue to evaluate and improve existing automated QC algorithms as well as develop new ones.

## 7. Summary

NDBC validates approximately six-million environmental buoy observations a year using automated near real time QC algorithms, and computer generated graphics and tools. The data from NDBC operated stations are archived at NCDC and NODC after a final monthly check. Scientists and technical support staff continue to work on improving the way data are quality controlled.

## References

- Conlee, D.T. and P.F. Moersdorf, 2004: The NWS Marine Observation Network: Coastal Marine Component of Multiple Observing Systems. *Ninth Symposium on Integrated Observing and Assimilation Systems for Atmosphere, Oceans, and Land Surface (IOAS- AOLS)*, Paper 7.1.
- Conlee, D.T. and P.F. Moersdorf, 2006: IOOS Backbone Expansion Efforts by NOAA's National Data Buoy Center. *Tenth Symposium on Integrated Observing and Assimilation Systems for Atmosphere, Oceans, and Land Surface (IOAS-AOLS)*, Paper 5.3.
- National Data Buoy Center, 2003: *Handbook of the Automated Quality Control Checks and Procedures of the National Data Buoy Center*.
- Gilhousen, D.B., 1998: Improved Real Time Quality Control of NDBC Measurements. *Preprints of the 10<sup>th</sup> Symposium of Meteorological Observations and Instruments*, Phoenix, AZ 363-366.

## THE ECOTRENDS WEB PORTAL: AN ARCHITECTURE FOR DATA DISCOVERY AND EXPLORATION

Mark Servilla<sup>1</sup>, Duane Costa<sup>1</sup>, Christine Laney<sup>2</sup>, Inigo San Gil<sup>1</sup>, and James Brunt<sup>1</sup>

<sup>1</sup>LTER Network Office, Department of Biology, MSC03 2020, 1 University of New Mexico, Albuquerque, NM 87131-0001

<sup>2</sup>Jornada Basin LTER, Box 30003, MSC 3JER, NMSU, Las Cruces, NM 88003-0003

### Abstract

The EcoTrends Project began in 2004 to promote and enable the use of long-term data to better understand processes within the Earth's ecosystems. Collected by local, state and federal agencies and institutions, these data are quality checked and then simplified into a common data format, called "derived data", for the EcoTrends database. A large-format book containing numerous time-series plots, along with vignettes of the derived data, will be published in 2008. Foresight by project coordinators realized the importance of also providing these data via the World Wide Web. A web-based portal would make possible the use of on-line tools that would streamline the discovery and exploration of these data, while at the same time facilitate adding new time-series data to the growing EcoTrends database. The EcoTrends Web Portal is now at its first milestone in production and includes features for data discovery and access by registered users, plotting both derived and smoothed data values, downloading both summary and statistically annotated data in comma delimited and HTML formats, and saving markers to high-value data in an on-line store that simplifies future access. The EcoTrends Web Portal utilizes the Provenance Aware Synthesis Tracking Architecture framework being developed by the LTER Network Office. This framework combines community developed open source tools, such as Metacat and Ecological Metadata Language, into a data warehouse workflow system that, when fully operational, will automate the extraction and uploading of site-based data into a permanent and persistent archive that can be utilized by synthesis projects.

**Keywords:** cyberinfrastructure, synthesis, web portal, long-term ecological research

### 1 Introduction

Research from around the globe is massing large volumes of data that span extended periods of time and come from a variety of different ecosystem settings. These long-term data are the focus of the EcoTrends Project, which began in 2004 as an informal discussion of how to promote such observations for use in broad-scale and significant synthesis projects. To date, 50 research sites, mostly from within the United States, now participate and contribute data to the EcoTrends Project (Peters and Laney 2006, Laney and Peters 2006), including Long-Term Ecological Research Network (LTER) sites, as well as sites supported by the U.S. Department of Agriculture (Agriculture Research Service and Forest Service), the U.S. Geological Survey, the U.S. Department of Energy, and other state institutions.

The contribution of site-collected data to the EcoTrends Project involves considerable management. All data are quality checked for accuracy and completeness, organized into a common data format, called "derived data", and then loaded into the EcoTrends database for use in community research and synthesis projects. The initial set of data are to be published in 2008 as a compendium of plots and vignettes in a large-format book that illustrates significant time-related trends of the derived data.

Project organizers decided that, in addition to the book, all derived data would be made available on the World Wide Web through a web-based portal application called the "EcoTrends

Web Portal” (<http://www.EcoTrends.info>). The EcoTrends Web Portal is a powerful resource for the community, providing a single point of access to an ever-growing set of environmental and ecological data that is organized in a common format, along with a set of tools for simple and quick discovery and visualization of temporal trends inherent within the data. This portal will accompany the publication and later succeed the book as additional data and new sites join the project. We anticipate over 20 thousand data sets will be available through the EcoTrends Web Portal when it is put into general production.

The following paper provides an overview of the EcoTrends Web Portal, including its architectural design and a discussion of salient features, as it meets its first operational milestone in early 2008.

## **2 Portal Architecture**

The EcoTrends Web Portal is designed with two functional goals in mind. First, the portal must manage existing data and support the addition of new data collected from sites that are already participating in the EcoTrends Project, as well as simplify the introduction of new sites and their data. Second, the portal must streamline access to data for users by providing “smart” data discovery and exploratory tools, including functions to quickly plot temporal trends of one or more data sets. The first goal is addressed by using the Provenance Aware Synthesis Tracking Architecture (PASTA) framework (Servilla et al. 2006) as the portal’s architectural foundation. The second goal is achieved through extensions of the current LTER Data Catalog (<http://metacat.lternet.edu>) data discovery interface.

### **2.1 PASTA**

The underlying design of the EcoTrends Web Portal is based on the PASTA framework that is being developed by the LTER Network Office as part of the nascent LTER Network Information System. This modular framework (Figure 1) is a design pattern for automating data collection from spatially separated locations into a centralized and persistent archive, which can be used as a data resource for further synthesis. To date, the framework has only been implemented within a development environment for testing purposes. The framework utilizes community developed tools, such as Metacat (a schema-independent XML database) (Berkley et al. 2001, Jones et al. 2001), Ecological Metadata Language (EML) (Nottrott et al. 1999, McCartney and Jones 2002, Fegraus et al. 2005), and the Data Manager library (Java functions within the EML distribution for extracting and loading tabular data described by EML) (O’Brien and Burt 2007), in addition to the open source PostgreSQL relational database management system and programmable interfaces (e.g., Java Servlets) for discovering and accessing archived or processed data. All software components developed as part of PASTA, including those of the EcoTrends Web Portal, are available as open source under the GNU General Public License version 2 (GPL2) through the LTER Network Concurrent Versions System (<http://cvs.lternet.edu>).

Data management begins when new “Source” data are added at a site and the corresponding EML document is updated and harvested into a local “Metacat” database. If the EML document identifier is registered in the “Dataset Registry”, an update to the EML document will trigger an automated data extraction event by the “Parser-Loader” module by using functions of the Data Manager library. The new data are added to the local “Cache” archive, which is available to the “Workflow Engine” for further processing. The “Workflow Engine” represents any transformation process that is necessary to generate derived data products

from the original source data. Data output from the “Workflow Engine” is stored in the “Derived Data” database, and metadata, as EML, is harvested back into the local “Metacat” database. External applications, such as web-based interfaces, are able to access derived data products by dereferencing links within EML documents discovered through Metacat’s client interface. New participating sites only need to commit to the metadata harvest process and have their EML document registered in the “Dataset Registry”.

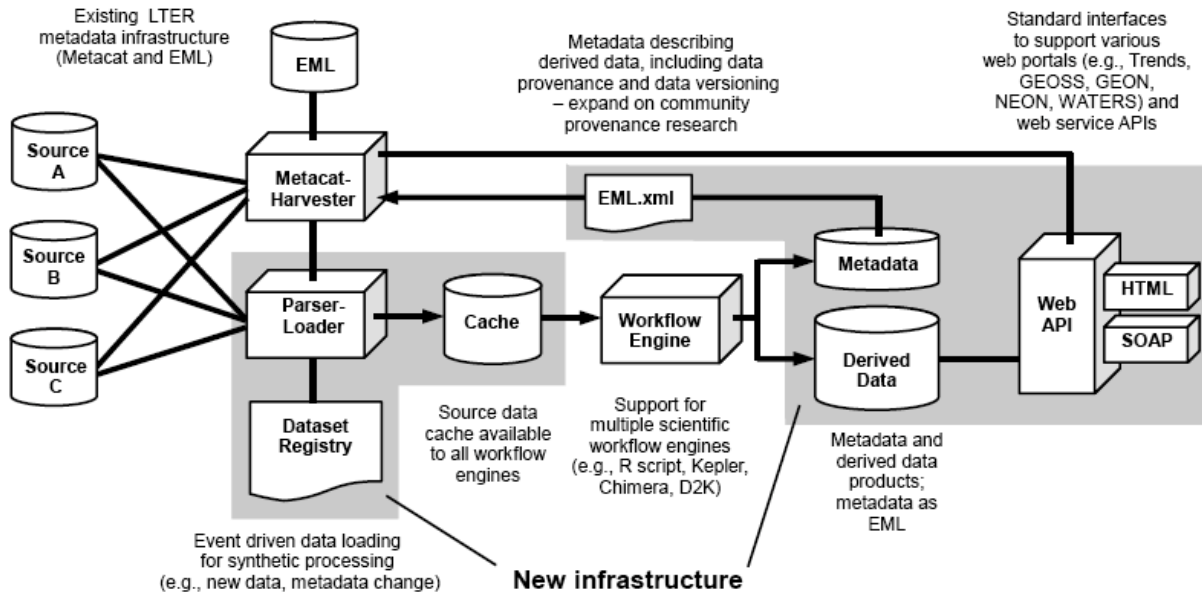


Figure 4 - Major components of the PASTA framework.

Although the PASTA framework is still under development, the EcoTrends Web Portal takes advantage of modules that support derived data products, including Metacat, EML, and a Java Servlet-based web application for data discovery and exploration.

## 2.2 Data Discovery and Exploration

Data discovery and exploration in the EcoTrends Web Portal starts with the search for derived data products by identifying relevant EML documents contained in the Metacat database. This process is achieved through either a “browse catalog” approach that uses a vocabulary of predefined terms as search criteria or a form-based approach that allows the user to select specific search criteria.

The “browse catalog” is separated into either “topic” or “site” sections, and are presented through separate web pages. The “topic” page is partitioned into biogeochemistry, biotic structure and disturbances, climate and physical variability, and human population and economy categories, each with a unique vocabulary of search terms. The “site” page lists each participating site in alphabetical order and uses the site name as the search term. Documents that are identified by the search are indexed into a list that is associated with each term and are made available through a single link from the corresponding web page.

In contrast to the “browse catalog”, the form-based approach supports both a simple “key-word” search page that allows a user to search on phrases containing one or more words and a complex, multi-field page for fine-tuning search criteria. The multi-field page allows the user to select from a combination of (1) the participating site name, (2) the data variable being

measured, (3) the temporal frequency of the derived data, (4) temporal coverage of the derived data, and or (5) the spatial bounding coordinates of the site where the data were originally collected as the criteria to search for EML documents. The site and variable names are displayed in a fixed “dropdown” list that is generated only when new sites or new data variables are added to the EcoTrends Project. Temporal coverage values are manually entered into a form field and are matched against the corresponding categories within the metadata. Spatial searches are matched against bounding coordinates also documented within the metadata. The spatial search tool provides form fields for manually entering bounding latitude and longitude coordinates or a Google map interface (<http://code.google.com/apis/maps>) that can be interactively adjusted to a visual bounding area.

Data discovery through either approach has the same effect. Search results are displayed as descriptive metadata in a table (Figure 2) for each identified data set, including the name of the participating site and data collection station, the “topic” category, the variable name, the temporal frequency of the derived data, and a set of “tool” icons. The “tool” icons provide the user with a set of functions for saving the selected data set in the local “My Data Store”, viewing detailed information about the data set, downloading the data in a “CSV” format, or viewing a time-series plot of the data. In addition, up to four data sets can be plotted together in a single plot (Figure 3). Users have the option to display only the data points, lines between data points, a line that is computed from a moving average of the data points, or any of the three options together. The “My Data Store” is a portal feature that lets users save data sets that are selected from the search results table. The same “tool” icons and plotting capabilities are available to all saved data sets in the “My Data Store”. The details of a data set can be viewed in a separate web page, which provides access to data in a “CSV” or “HTML” format and to metadata in the native XML structure of the EML document or in a nicely formatted “HTML” presentation that is generated by the Metacat. Access to all versions of data and metadata and the same plotting capabilities available from the search results table is also found on this “details” page.

### **3 Discussion and Conclusion**

The World Wide Web provides a wonderful abstraction through its server-based applications such that web content can be seamlessly updated without effort for the consumer of that content. The EcoTrends Web Portal is no exception. Its primary goal is to provide access to updated and newly derived data today and into the future. Most of the long-term environmental and ecological data being collected and processed for the EcoTrends Project are part of ongoing research. These projects are continually adding new data to their holdings and provide the EcoTrends Web Portal a rich resource from which derived data products can be generated and new synthesis research may take form. Providing the most up-to-date data to the community is paramount to the success of the EcoTrends Web Portal. At present, site-based data must be manually loaded into the “Cache” database for use in generating derived data products. By using the PASTA framework as its core cyberinfrastructure, the EcoTrends Web Portal is strategically positioned to automate its site-based data loading and generation of derived data products when future standards enable widespread and rich metadata in EML. In addition, PASTA’s use of open source components ensures synergy with the broader ecoinformatics community into the future.

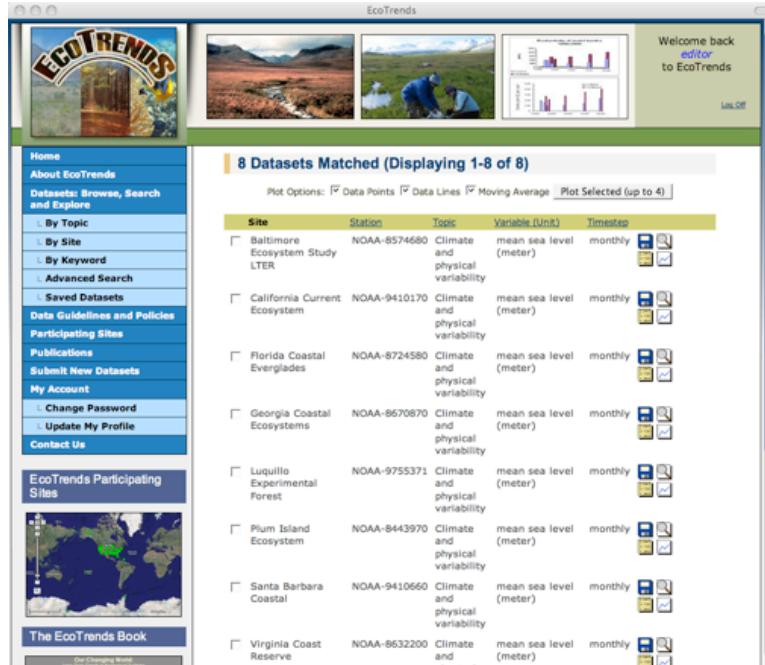


Figure 5 - Display of the search results table from the EcoTrends Web Portal.

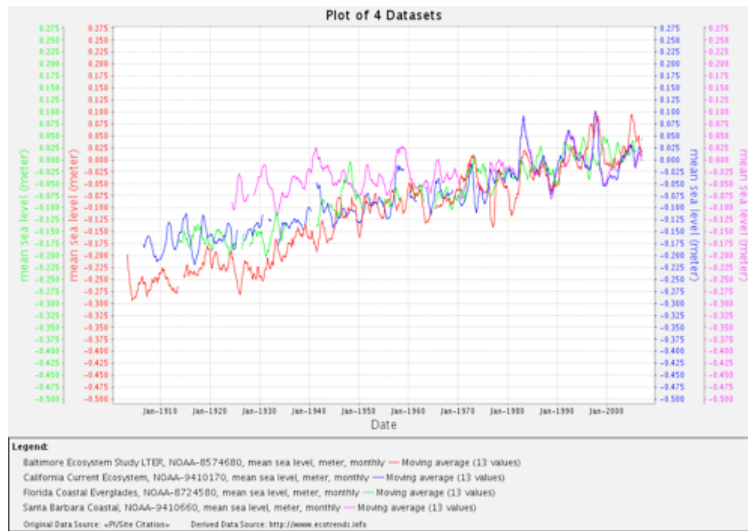


Figure 6 - Time-series plot of derived measurements from four different sites.

An equally important goal for the EcoTrends Web Portal is the ability for users to quickly discover meaningful data. The two different portal approaches for data discovery streamlines the “exploration” process of finding the right data. The “Browse Catalog” satisfies most user needs by directing their search effort to specific categories. The “pre-searched” terms provide nearly instant display of information on available data and are refreshed nightly, minimizing the probability of missing data that was recently added. When the “Browse Catalog” is too generalized, the form-based search allows the user to “fine-tune” search criteria for better control of search fidelity. To reduce the time of repeated searches, the portal’s “My Data Store” can save key data sets for future use. Access to previous versions of both data and metadata is also a noteworthy feature of the portal, since reproducibility of synthetic data products that were based



on earlier versions of any derived data may be required to verify peer-reviewed studies. Perhaps, the most important feature to identify meaningful data available through the EcoTrends Web Portal is interactive plotting. Real-time plots of multiple data sets can be used to visualize and quickly interpret temporal trends of derived data, thereby focusing effort on data that is significant to the end user.

### **Acknowledgments**

The authors of this paper would like to thank Dr. Debra Peters, EcoTrends Project director, and the science members of the EcoTrends Project Committee for their thoughtful review and evaluation of the EcoTrends Web Portal. We gratefully thank the technical members of the EcoTrends Project Committee, Don Henshaw, Ken Ramsey, Mark Schildhauer, Wade Sheldon, and Marshall White, for their countless hours given to discussions on the best design of EcoTrends Web Portal. The development of the EcoTrends Web Portal is supported by the National Science Foundation under Grant number DEB-0080412 and Cooperative Agreement DEB-0236154.

### **References**

- Berkley, C., M. Jones, J. Bojilova, and D. Higgins, 2001. Metacat: A schema-independent XML database system. 13th Intl. Conference on Scientific and Statistical Database Management: 171.
- Fegraus, E.H., S. Andelman, M.B. Jones, and M. Schildhauer, 2005. Maximizing the value of ecological data with structured metadata: An introduction to Ecological Metadata Language (EML) and principles for metadata creation. *Bulletin of Ecological Society of America*. 86(3): 158-168. doi: 10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2.
- Jones, M.B., C. Berkley, J. Bojilova, M. Schildhauer, 2001. Managing scientific metadata, *IEEE Internet Computing* 5(5): 59-68.
- Laney, C.M. and D.P.C. Peters, 2006. EcoTrends in long-term ecological data: A collaborative synthesis project, introduction and update. *ILTER DataBits Spring 2006*: (<http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/06spring/>)
- Nottrott, R., M.B. Jones, and M. Schildhauer, 1999. Using XML-structured metadata to automate quality assurance processing for ecological data. *Proceedings of the Third IEEE Computer Society Metadata Conference*. Bethesda, MD, April 6-7, 1999.
- McCartney, P. and M. Jones, 2002. Using XML-encoded metadata as a basis for advanced information systems for ecological research. *Proceedings of the 6th World Multiconference Systemics, Cybernetics and Informatics, (7)*, International Institute for Informatics and Systematics, 2002, pp. 379-384.
- O'Brien, M. and C. Burt, 2007. A query interface for EML data tables. *ILTER DataBits Spring 2007*: (<http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/07spring/>)
- Peters, D. and C. Laney, 2006. Trends in Long-Term Ecological Research projects. *Jornada Trails* 10(1): 2.
- Servilla, M., J. Brunt, I. San Gil, and D. Costa, 2006. PASTA: A network-level architecture design for generating synthetic data products in the LTER Network. *ILTER DataBits Fall 2006*: (<http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/06fall/>).

## **DYNAMIC, RULE-BASED QUALITY CONTROL FRAMEWORK FOR REAL-TIME SENSOR DATA**

**Wade M. Sheldon**

Department of Marine Sciences, University of Georgia, Athens, Georgia, USA

### **Abstract**

The volume of monitoring data that can be acquired and managed by Long Term Ecological Research sites and environmental observatories has increased exponentially over time, thanks to advances in sensor technology and computing power combined with steady decreases in data storage costs. New directions in environmental monitoring, such as sensor networks and instrumented platforms with real-time data telemetry, are raising the bar even higher. Quality control is often a major challenge with real-time data, though, due to poor scalability of traditional software tools, approaches and analysis methods. Software developed at the Georgia Coastal Ecosystems Long Term Ecological Research Site (GCE Data Toolbox for MATLAB) has proven very effective for quality control of both real-time and legacy data, as well as interactive analysis during post processing and synthesis. This paper describes the design and operation of the dynamic, rule-based quality control framework provided by this software, and presents quantitative performance data that demonstrate these tools can efficiently perform quality analysis on million-record data sets using commodity computer hardware.

**Keywords:** quality control, statistical analysis, real-time data, sensor, MATLAB

### **1. Introduction**

Quality control is a critical component of environmental data management, particularly for data collected by autonomous sensors. Many factors can affect the quality of sensor data, including calibration drift, biological fouling, electrical noise during data transmission, and mechanical interference from other instruments or mounting hardware (Gentili et al, 2004; Magnaterra et al., 2004). These problems lead to data contamination that can profoundly affect data analysis and skew interpretation. Traditionally, quality control of environmental data has been conducted by visually inspecting or plotting data values and performing detailed statistical analyses (e.g. distribution tests, outlier tests) using specialized software (Edwards, 2000). However, the sheer number of parameters and volume of data generated by modern sensors and sensor networks often precludes this approach. Consequently, some monitoring programs (e.g. U.S.G.S. National Water Information System) report provisional near-real-time data with no or minimal quality control processing, then release reviewed, derived data products at a later time.

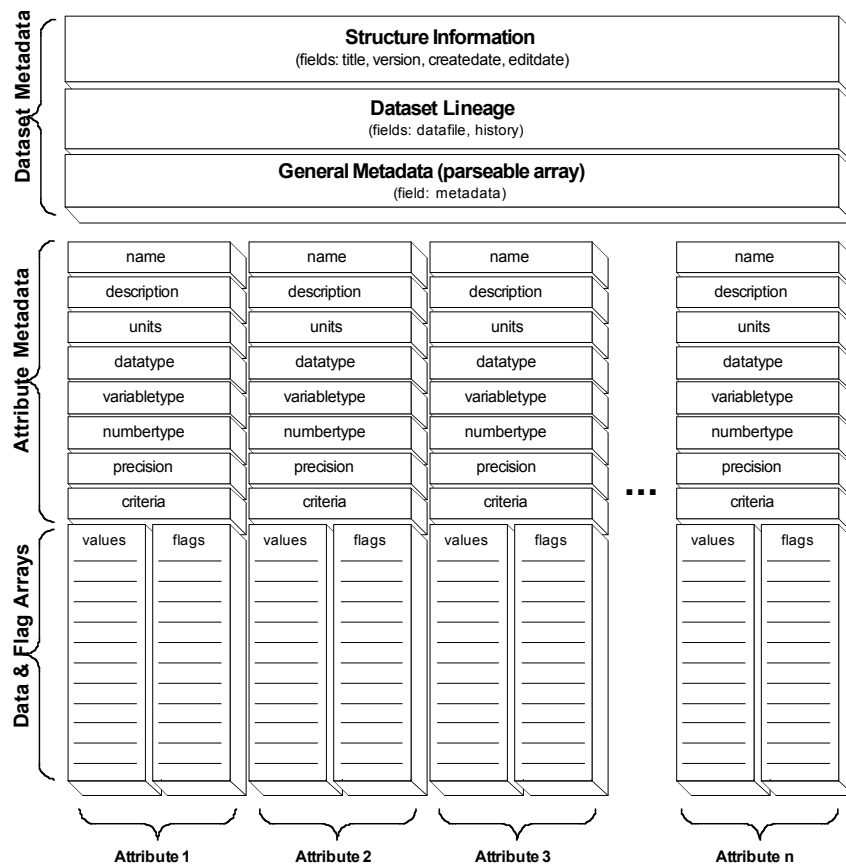
Software developed at the Georgia Coastal Ecosystems Long Term Ecological Research Site (<http://gce-lter.marsci.uga.edu>), the GCE Data Toolbox for MATLAB, includes a dynamic, metadata-based quality control framework that has proven useful for analysis of real-time sensor data, as well as non-real-time and legacy data sets. Although other metadata-based quality control processing approaches have been advanced (Nottrott et al., 1999), this software provides a fully integrated, extensible solution that supports both automated and interactive analysis within a seamless software environment. An unlimited number of quality control “rules” can be defined for each parameter in a data set, and rules are evaluated automatically whenever data are imported or revised to generate alphanumeric “flags” that are intrinsically managed along with

the data values they qualify. This paper describes the design and operation of the quality control framework provided by this software, and its potential use for high volume sensor data sets.

## 2. Methods and Techniques

The GCE Data Toolbox software package (Sheldon, 2002) was developed using the MATLAB<sup>®</sup> technical programming language (The MathWorks, <http://www.mathworks.com>). MATLAB was selected because of its prevalence in environmental science and engineering as well as its excellent support for large data sets (limited only by computer memory) and code portability across Windows, UNIX and Macintosh computer platforms. MATLAB is also a dynamically-typed, interpreted language, making it well suited for rapid software development, testing and deployment (Prechelt, 2000).

The GCE Data Toolbox intrinsically supports data quality control at all levels, starting with the underlying data model (fig. 1). Data sets are managed by toolbox programs as



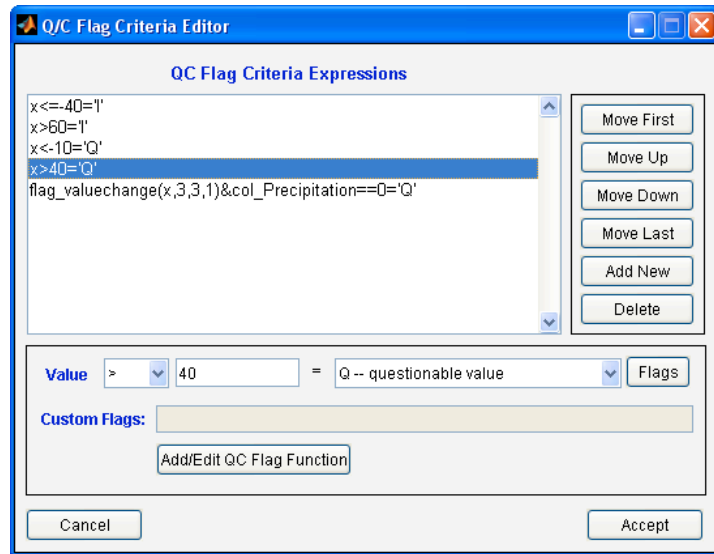
**Figure 1.** Conceptual model of the GCE Data Structure (version 1.1, 29-Mar-2001), illustrating the types and cardinalities of metadata fields, data arrays and quality control flag arrays. Structures are created and managed using the GCE Data Toolbox, a MATLAB software library for metadata-based analysis, visualization and management of ecological data sets.

MATLAB structure arrays, with dedicated fields for data set metadata and lineage, and repeating groups of attribute metadata fields, data arrays, and quality control flag arrays, which are managed collectively as data set attributes. Correspondence of data values and flags is

maintained across attributes throughout all data manipulation operations (e.g. sorting, filtering, joins, unions) similarly to tuples in a relational database model. Flags are stored as single-character alphanumeric codes, which are defined in the data set metadata. An empty string denotes absence of a flag, and multiple flags can be assigned to a single data value.

Quality control rules are defined using the syntax: [expression]='[flag code]', where [expression] is any MATLAB statement that returns a logical array of 1's and 0's, and [flag code] is the alphanumeric character to assign to values matching the criteria (i.e. [expression] evaluates as 1). Data columns are referenced in rules using "x" to represent the current column, or "col\_[column name]" to reference any column in the data set by name. For example, the rule "x<0='Q'" assigns "Q" flags to any negative values in the corresponding data column, and "col\_Dry\_Weight<(col\_Wet\_Weight-col\_Ash\_Weight)\*0.90='I'" (in column Dry\_Weight) assigns "I" flags to any values of Dry\_Weight that are less than 90% of the difference between Wet\_Weight and Ash\_Weight. Compound rules can be defined by separating multiple rule expressions with semicolons (e.g. "x<0='I';x>100='I'; x<20='Q';x>80='Q'"). Rule statements are stored in the "criteria" metadata field for each attribute, and can be defined in advance using metadata templates or created and edited interactively using a GUI form (fig. 2).

Rules can be defined to perform a wide variety of quality control analyses based on numeric, text and statistical comparisons using this simple syntax (Table 1). The default framework can also be extended simply by adding custom MATLAB functions to the toolbox directory and referencing these functions in quality control rules. Custom functions can be written to retrieve reference data from a file system, database query or web service then run complex algorithms or models implemented using MATLAB or another programming language (e.g. Java or FORTRAN), so the potential scope of rules is unlimited. It should be noted, though, that although powerful, this feature does represent a potential security risk. Overt system calls in rules are blocked and error handling routines prevent syntax and run-time errors from halting the program or corrupting data, but a malicious user could inject calls to external functions capable of altering data or launching attacks, so access to data set metadata and templates should be controlled in a network setting.



**Figure 2.** Graphical quality control rule editor form in the GCE Data Toolbox for MATLAB

**Table 1.** Representative quality control operations, rule types and syntax. Note that “x” symbols in rule criteria are aliases for values in the corresponding data column, and that the “col\_” prefix denotes values from any data set column referenced by name (including the column containing the rule).

Operation	Quality Control Goal	Rule Type	Example Syntax
Range check	Confirm values are within range of the sensor/parameter	numeric conditional	<code>x&lt;0='I';x&gt;100='I'</code> -- assigns I flags to negative values and values over 100
Consistency check	Confirm values are consistent with other measured parameters or historic maxima/minima	multi-column conditional	<code>col_DOC_Conc&gt;col_TOC_Conc='I'</code> -- assigns I flags to DOC concentrations that exceed total organic carbon concentration
		statistical expression	<code>x&gt;mean(x)+4*std(x)='Q'</code> -- assigns Q flags to values more than 4 standard deviations above the column mean
Vocabulary check	Confirm values conform to a controlled vocabulary (e.g. standard code list)	custom function	<code>flag_notinlist(x,'A1,A2,A3,A4')='Q'</code> -- assigns Q flags to values not in the specified list (or referenced data set)
Dependency check	Confirm measurements were recorded under suitable conditions, based on other parameter observations	multi-column conditional	<code>col_Depth&lt;0.1='Q'</code> -- in column Salinity; assigns Q flags to values recorded when instrument depth was < 0.1m, indicating water emergence
Pattern check	Confirm values do not exhibit temporal or spatial patterns that indicate sensor failure or data contamination	custom function	<code>flag_percentchange(x,25,25,3)='S'</code> -- assigns S flags to values that are >25% above or below the mean of the preceding 3 values
Reference data check	Confirm values agree with prior recorded values or reference values	custom function	<code>flag_locationcoords(x,col_Lon,col_Lat,0.2)='I'</code> -- assigns I flags to location codes that differ by more than 0.2km from the registered coordinates, based on corresponding Latitude and Longitude values

Quality control rules are automatically evaluated to assign or clear flags whenever data values are entered, imported or edited (or the rules themselves are revised) using toolbox functions. Flags can also be assigned manually with the mouse on data plots or using a spreadsheet-like GUI editor to augment or revise rule-based flag assignments. Additionally, flags can be parsed from text attributes in data sets, allowing flags assigned by other data management systems to be imported into the toolbox framework. When flags are defined manually or imported, the token “manual” is added to the corresponding “criteria” attribute metadata field. This token locks flags for the data column so manually-assigned flags are not subsequently overridden by automatic rules. Removing the manual token restores automatic flag evaluation.

Quality control rules and flags are constitutive components of the GCE Data Toolbox data model, so most toolbox functions provide explicit options for handling flagged values during post processing and analysis. For example, flags can be displayed, ignored or flagged-values removed when data are plotted, and statistics reports can be generated with and without flagged values. Data export functions provide various options for formatting flags in delimited text and MATLAB files to support other programs and standards, and data integration tools (e.g. union and join functions) provide options for automatically locking flags to prevent inappropriate application of criteria after multiple data sets are combined. In addition, data aggregation, date/time re-sampling, and binning tools optionally create quality control rules for all derived

data columns based on the number or percentage of flagged and missing values in each respective group, date/time interval or bin to provide quality control for derived data products.

In order to test the performance of the GCE Data Toolbox for quality control analysis of high volume sensor array data sets, a 1,000,000 record by 48 column time series data set was compiled from various sources (i.e. equivalent to one year of observations at 30 sec frequency). Three to six quality control rules were defined for each column, including numeric range checks, statistical consistency checks, and multi-column dependency checks, resulting in flags being assigned to 0-14% (mean 4.3%) of values. The test data table was subset into 12, 24 and 48 column tables of varying length, and the time required to evaluate all rules and manage assigned flags using the “dataflag” function was evaluated for two different versions of MATLAB (release 2007b and release 13a/version 6.51) on a Dell® computer with an Intel® Core Duo™ T2500 processor (2.0 GHz clock speed) and 1 GB RAM.

### 3. Results and Discussion

In performance testing, quality control rule evaluation time varied linearly with number of records and number of parameters (Table. 2), with both slopes near unity. These results and additional trials with larger tables on computers with up to 4 GB of system RAM indicate that algorithm execution time is directly proportional to table size for a given rule set, software and hardware configuration. Evaluating rules for the complete 1 million record data set required <42 sec on the test hardware, and a 100,000 record data set required <4 sec, indicating that interactive quality control analysis of typical sensor data sets is very practical.

**Table 2.** Quality control rule evaluation time in seconds versus data set table size. Timings were evaluated using MATLAB releases 13a (R13a) and 2007b (R2007b).

Data Set Records	12 Parameters (30 rules)		24 Parameters (60 rules)		48 Parameters (120 rules)	
	R13a	R2007b	R13a	R2007b	R13a	R2007b
10,000	0.09	0.08	0.19	0.11	0.39	0.22
50,000	0.45	0.25	0.97	0.51	1.81	1.02
100,000	0.97	0.58	1.97	1.16	3.94	2.36
200,000	1.98	1.17	3.97	2.36	7.95	4.77
400,000	4.03	2.41	8.06	4.86	16.20	9.77
600,000	6.11	3.67	12.25	7.36	24.59	14.88
800,000	8.19	4.91	16.39	9.89	32.84	19.89
1,000,000	10.25	6.14	20.56	12.38	41.08	26.58

The GCE Data Toolbox was developed as a comprehensive data processing solution for GCE information management staff, but it has proven useful beyond this scope. Several GCE investigators and many graduate students have used the toolbox to analyze core GCE data as well as their own data, and the toolbox is used in Marine Science methods classes at UGA. Since its initial release in 2002, over 2800 web visitors not affiliated with the GCE project have downloaded the toolbox for a wide-ranging set of applications (e.g. hydrological data analysis, quality control of U.S.A.F. test flight data, U.S.G.S. and LTER ClimDB data mining). Although formal usability testing has not been performed to date, various enhancements requested by users have been implemented and user feedback on functionality has been uniformly positive.

#### 4. Conclusions

The GCE Data Toolbox is well suited for processing high volume, real-time sensor array data and performing quality control analysis. Metadata templates containing detailed attribute descriptors and quality control rules can be defined using GUI forms for each sensor platform, and then applied automatically to validate and flag data values when raw data are imported from data loggers, text files, or database queries. Data summaries and plots can be generated to review the quality control analysis results, then flag rules and flag assignments can be refined as necessary. Derived data products (containing additional quality control rules) can then be generated for distribution or further analysis, preserving information about the quality and completeness of the source data in the data set metadata, derived attributes, and quality control rules. This workflow can then be automated for routine data acquisition and analysis.

The performance results reported above indicate that the flag evaluation algorithms are suitably efficient for processing million-record data sets in real time on commodity computer hardware, with maximum data set size only limited by available system RAM. Multiple instances of the GCE Data Toolbox can also be run on a single computer to simultaneously process multiple data streams, minimizing the MATLAB licenses required to use the software.

A compiled version of the toolbox is publicly available online (Sheldon, 2002), and source code is available on request for evaluation and end-user customization. Offers to collaborate on future toolbox development are also welcome.

#### Acknowledgments

This material is based upon work supported by the National Science Foundation under grant numbers OCE-9982133 and OCE-0620959.

#### References

- Edwards, D. 2000. Data Quality Assurance. Pg. 70-91 in: Michener, W.K. and Brunt, J.W., eds. Ecological Data - Design, Management and Processing. Blackwell Scientific, Oxford.
- Gentili, S., Magnaterra, L. and Passerini, G. 2004. An introduction to the statistical filling of environmental data time series. Pg. 1-27 in: Latini, G. and Passerini, G., eds. Handling Missing Data: Applications to Environmental Analysis. WIT Press, Boston.
- Magnaterra, L., Passerini, G. and Tascini, S. 2004. Data validation and data gaps in environmental time series. Pg. 29-89 in: Latini, G. and Passerini, G., eds. Handling Missing Data: Applications to Environmental Analysis. WIT Press, Boston.
- Nottrott, R., Jones, M.B. and Schildhauer, M.P. 1999. Using XML-Structured Metadata to automate quality assurance processing for ecological data. Proceedings of the Third IEEE Computer Society Metadata Conference. IEEE. Bethesda, MD.
- Prechelt, L. 2000. An Empirical Comparison of Seven Programming Languages. IEEE Computer, 33(10):23-29.
- Sheldon, W.M. 2002. GCE Data Toolbox for MATLAB – Software tools for metadata-based analysis, visualization and transformation of ecological data sets. ([http://gce-lter.marsci.uga.edu/public/im/tools/data\\_toolbox.htm](http://gce-lter.marsci.uga.edu/public/im/tools/data_toolbox.htm))

## **STREAMING SENSOR DATA FROM SPACE: ACQUIRING AND MANAGING DIRECT BROADCAST SATELLITE DATA FOR SITES OF THE LONG TERM ECOLOGICAL RESEARCH NETWORK**

**John Vande Castle and Mark Servilla**

Long Term Ecological Research Network Office, Department of Biology, University of New Mexico MSC03 2020, Albuquerque, NM 87131-0001

### **Abstract**

Satellite sensor data provide important information over extensive areas but are usually not considered in the realm of real-time sensor data since most satellite information is used after post-processing, from archive centers weeks, months or longer after initially acquired. However, with the advance in multi-core processor speed and data storage technologies, direct broadcast data transmission from environmental satellite systems can now be accessed on a near real-time basis. Processing the raw satellite data, particularly from the MODIS sensor of the Terra and Aqua spacecraft and the AVHRR sensor of the NOAA series spacecraft into usable data products is now achieved within minutes of a satellite overpass. Current direct readout processing systems are able to downlink the raw sensor data, store and reprocess the data in near real-time. These systems can be configured for automated acquisition and processing to provide standardized data products within minutes of acquisition. This provides an important data source for ecological disturbance events including hurricanes, storms, flood, fire and other applications.

Collaboration between the Center for Rapid Environmental Assessment and Terrain Evaluation and the Long Term Ecological Research Network Office acquires processes and distributes these data for use by scientists and researchers. Data products produced from the raw sensor data are transferred to the data storage systems of web servers in near real-time for access. Automated generation and harvesting of metadata from the data products are updated to servers as the data are produced. This paper describes the flow of raw data from the satellite sensor to standard products for the near real-time archive and the generation of metadata to provide for search capability of the data.

**Keywords:** Satellite data acquisition, automated data processing, MODIS

### **1. Introduction:**

Remote sensing data from satellite sensor systems represent a standardized source of environmental information that can cover areas from meters to kilometers. However these data historically required extensive post-processing after they were downlinked and transferred to government or commercial data processing facilities. This delay in processing and delivery contributed to the general use of remote sensing data primarily for ad-hoc analysis. Direct broadcast capabilities of some environmental satellites, particularly from the Advanced High Resolution Radiometer (AVHRR) sensor of the NOAA series of polar orbiting satellites and more recently, the Moderate Resolution Imaging Spectroradiometer (MODIS) sensor of the NASA Terra and Aqua satellites permit direct reception of data from these sensors. These data have been available from the NOAA spacecraft since the mid 1980's and NASA satellites shortly after the launch of Terra-1 in December of 1999. However computational power and software availability were a hindrance to rapid use of data from sensors on the spacecraft.



Until recently, the high data rates and extensive processing still meant these data were not available until days, weeks or longer after reception. However, with the advances in processing speeds and data storage capacities, delays from post-processing of satellite data is no longer an issue in the use of the data. With the previous development of computer cluster and RAID technologies, satellite data processing resulted in final product generation within an hour or so of data reception. However, with advances in multi-core processors, large physical memory and local disk storage exceeding 1Tb, satellite data processing and product generation can now occur in near real-time, with final product generation within 10 minutes of data reception (Prasad, 2008). In addition, the capabilities of newer automated systems allow for data acquisition of specific regions within the programming framework of the direct readout systems. The data can be acquired, processed, stored and distributed in a completely automated operation.

The rapid availability of direct broadcast environmental satellite data from these space-borne sensors represents a new form of real-time streaming sensor technology. The significance of the direct broadcast data is the capability to provide information concurrent with disturbance events such as elevated heat periods, storms or fires. The data become available immediately following a satellite overpass, and are data from a single overpass rather than archived composite data from multiple satellite passes over a time frame of two or more weeks. Data from direct broadcast systems are not meant to replace more conventional sources of data products such as from NASA Distribute Active Archive Centers, (DAAC), but rather augment these data sources with near real-time data in standardized and simple data formats. For instance, MODIS subset data can be acquired from the Oak Ridge National Laboratory DAAC ([http://www.modis.ornl.gov/modis/modis\\_subsets1.cfm](http://www.modis.ornl.gov/modis/modis_subsets1.cfm)), but only with the current processing delays of days to weeks or longer. Data can also be acquired from NASA rapid response centers (i.e. <http://rapidfire.sci.gsfc.nasa.gov>) but these data are generally available only as image data without calibration information needed to fully interpret the data.

As with the advances in land-based sensor technologies and data transmission through wireless technologies, the challenges permitted by real-time processing and data storage still lie in **using** these data in real-time situations. The near real-time availability of current data provides the ability to use this information in predictive, rather than just descriptive research.

A collaboration between the Long Term Ecological Research (LTER) Network Office of the LTER Network Program and the Center for Rapid Environmental Assessment and Terrain Evaluation (CREATE) at the University of New Mexico provides near real-time data products for sites of the LTER Network. The direct readout facilities at CREATE download direct broadcast AVHRR from the NOAA series of satellites as well as MODIS data from the NASA Terra and Aqua satellites covering most of the conterminous United States. Although AVHRR data including vegetation index and temperature products are produced by CREATE, the focus of near real-time data for LTER sites has been with the more advanced capabilities offered by the MODIS data. Initial acquisitions for LTER sites started in July of 2007 and the remaining sites were on-line by January of 2008. Currently 22 of the 26 sites are within view of the CREATE satellite antennae and efforts are underway to access direct broadcast data from other receiving stations to cover the rest of the LTER Network. Such Network-wide remote sensing acquisitions have been performed in the past (Vande Castle, 1998) and enabled cross-site comparisons (Riera et al., 1998). Although this work was with higher spatial resolution Landsat data, the higher temporal resolution MODIS data provide new opportunities for research based on their daily availability.

## **2. Satellite Data Acquisition and Processing:**

Reception of data from environmental satellite systems requires accurate satellite tracking and downlink of the data in real time. The current turn-key automated systems of CREATE acquire satellite orbit information from NASA and other internet servers to provide tracking information for automated data reception of satellite passes within view of the receiving station. This tracking information is linked to automated scheduling software to receive the direct broadcast satellite data on an unattended basis. The direct broadcast data of the MODIS sensor on the Terra and Aqua satellites is transmitted to direct readout stations on a X-band radio frequency of about 8.2 GHz with a compressed data rate of about 13Mbps. This data transmission is quit analogous to wireless sensor transmission from land-based sensors to internet servers except for the data rates and more complicated data formats. Reception of data from a 16 minute overpass typically generates about 1.5gb of raw compressed data.

The raw direct broadcast data are first preprocessed, involving data decompression and frame synchronization into standard raster format data of rows and columns representing standard imagery products. This produces standardized data that is known as a production data set or "PDS" file. All of these data are archived to a standard RAID system as part of the automated processing for backup or future reprocessing. Once generated, the PDS data files are ingested by a series of processing steps to produce final data products. This processing includes the addition of satellite ephemeris, or location data to provide spatial attribute data needed for further processing and production of the final data products. Through a number of automated processing steps, the raw data are used to produce a variety of standard data products such as corrected radiance, corrected reflectance, Normalized Difference Vegetation Index (NDVI), surface temperature and other data products in a standardized Geographic Information System (GIS) format. Producing these full-pass data products is the most time consuming part of the processing step, although production of the large number of subset products for LTER sites and other areas impacts the processing as well.

For LTER site data acquisition, scripts including site location data and projection information provide a template to extract site data from the full pass data scenes as part of the automated processing. This processing currently extracts MODIS data from the CREATE processing stream for 128km by 128 km regions encompassing LTER sites. Depending on the MODIS data product, spatial resolution varies from 250 meters for MODIS bands 1 and 2 and related products including normalized vegetation indexes (NDVI), 500 meters for MODIS bands 3-7 and related products and 1 kilometer resolution for other MODIS data. Larger areas are also generated for other projects and the spatial area for LTER sites could be increased based on feedback by LTER scientists. A focus of the CREATE processing is to produce MODIS products similar to the standard products distributed by the NASA archive centers except each product represents data from a single satellite overpass rather than aggregated data such as the NASA 16-day Composite NDVI data product. The processing for data products use the current NASA algorithms for processing MODIS data, modified for direct broadcast applications. These algorithms are based on the original algorithms developed by NASA researchers (e.g. Huete et al. 2002, Morisette et al., 2002).

The data processing for LTER sites include simple JPEG browse images meant primarily as a reference within a quick look gallery for screening of the data products. The primary focus of the processing is to produce standard data products such as NDVI, corrected reflectance, cloud masks, or thermal radiance products in a readily usable format. Although all data are initially generated in Hierarchical Data Format (HDF), all products are translated to GEOTiff format as

part of the processing stream. The GEOTiff data are designed for input to standard GIS and image processing software such as ENVI, or Erdas/Imagine.

The data products produced for each of the LTER sites differs for terrestrial and aquatic/coastal sites, with products such as NDVI, fire, and surface temperature produced for terrestrial sites and standard products such as sea surface temperature added to coastal sites. The methodology for producing a specific product is consistent across all sites for comparison of the data. In addition, all geospatial/raster information of similar products is identical for each LTER site so the products can be compared without reformatting. For example, map projection, coordinate information, and image structure is identical for all data products generated for an individual LTER site regardless of collection date and time. All data products are automatically transferred to RAID storage of a web server for data access and the browse images incorporated into an image gallery for viewing. Direct download of individual data products as well as ftp access to the complete archive is available on the web server.

To enable internet search access of the data, secondary metadata of the products is produced. The descriptive and accessible metadata for the individual data products are contained only within the header information of the self-describing GeoTiff files on the CREATE web server. Following previous work to generate metadata for LTER Landsat data in the LTER Network Office archive, all CREATE data products are processed to generate metadata documentation in the Ecological Metadata Language (EML) standard and make the EML available through the LTER Metacat Data Catalog to facilitate discovery by LTER Network scientists and the broader community. For these data, the EML metadata consist of standardized content that describe project information, contact information, methodology used to create the product, and the geospatial/raster information of the data product. A single EML document is produced for each product or dataset that is associated with each of the LTER sites rather than producing an EML document for every image product to remove the redundancy of products in data searches. New product metadata including temporal information are added to the corresponding EML document as new data enter the archive and the EML revision value is changed. This metadata consist of a collection date and data URL that explicitly reference each data product. The collection date, as part of the data URL string, updates the temporal coverage element of the EML document so the data will appear in any a search of a specific date or spatial range where they occur.

### **3. Future Directions:**

The acquisition of the real-time environmental satellite data represents a new way to look at information available from space-based sensor systems. For LTER research these data can be used immediately following disturbance events such as extended heat stress, storms or fires. A major challenge still lies in the use of the data in real-time situations. For example, a link between rodent population changes and increased vegetation after rainfall events has been documented through past NDVI data analysis (Yates et al., 2002), however archived AVHRR data were used in this analysis. MODIS time series data from NASA archives has also been used to describe seasonal changes in the vegetation signal of LTER sites (Vande Castle, 2003). However, new techniques for linking real-time processes to data are important for future research efforts.

Initial research in comparing data products such as vegetation index and pollen data or temperature data with health statistics including asthma and heat related hospitalization data are underway through other collaborations with CREATE. The near real-time availability of current

data provides the ability to use this information for predictive, rather than just descriptive, research. For the LTER MODIS data, new efforts are underway to revise all processing based on open source processing algorithms and techniques. This processing follows algorithms such as the International MODIS/AIRS Processing Package (Huang, 2004) and others currently in use at the University of Wisconsin Space Science and Engineering Center. New processing of data for LTER sites is also underway using an alpha version of the International Polar Orbiter Processing Package (IPOP, 2008). The software is specifically designed to run on standard off the shelf Linux multi-core computer hardware to process current MODIS data as well as data of future sensors such as the Visible/Infrared Imager/Radiometer Suite (VIIRS) sensor system which will be used on successors to the NASA Terra and Aqua spacecraft. Additional information and access to the CREATE data products for LTER sites can be found on the LTER remote sensing/GIS web page at: <http://www.lternet.edu/technology/ltergis/> or on the CREATE web site at <http://create.hpc.unm.edu/create/lter.php>

### References

- Huang, Hung Lung, I. Gumley, K. Strabala, J. Li, E. Weisz, T. Rink, K. Baggett, J. Davies, W. Smith and J. Dodge. 2004. The International MODIS/AIRS Processing Package. *Bulletin of the American Meteorological Society*, 85:2
- Huete, A., K. Didan, T. Miura et al., "Overview of the radiometric and biophysical performance of the MODIS vegetation indices," *Remote Sens. Environ.*, vol. 83, pp. 195- 213, 2002.
- IPOP, 2008. International Polar Orbiter Processor Package.  
<http://directreadout.sci.gsfc.nasa.gov/>
- Morissette, J.T., J. L Privette, et al., "A framework for the validation of MODIS Land products," *Remote Sens. Environ.*, vol. 83, pp. 77-96, 2002.
- Prasad, Kota 2008. Proceedings: NASA International EOS/NPP Direct Broadcast Meeting.  
[http://dbmeeting.sci.gsfc.nasa.gov/posters\\_presentations2008.cfm](http://dbmeeting.sci.gsfc.nasa.gov/posters_presentations2008.cfm)
- Riera, J, J. Magnuson, J.Vande Castle and M.MacKenzie 1998. Analysis of Large-Scale Spatial Heterogeneity in Vegetation Indices among North American Landscapes. *Ecosystems* 1:268-282
- Vande Castle, J.R. 1998. Remote sensing applications in ecosystem analysis. In: *Scale Issues in Ecology*. 13:271-288. D. Peterson and V.T. Parker Eds. Columbia University Press.
- Vande Castle, J.R. 2003. Vegetation Change Observations of Long-Term Ecological Research Sites Using Remote Sensing Data. Proceedings - 30<sup>th</sup> Symposium of the International Society on Remote Sensing of the Environment: TS-42.4
- Yates T.L., J.N. Mills, C.A. Parmenter, T.G. Ksiazek, R.R. Parmenter, J. Vande Castle, C.H. Calisher, S.T. Nichol, K.D. Abbott, J.C. Young, M.L. Morrison, B.J. Beaty, J.L. Dunnun, R.J. Baker, J. Salazar-Bravo and C.J. Peters. 2002 – Ecology and Evolutionary History of and Emergent Disease: Hantavirus Pulmonary Syndrome. *Bioscience* 52:11;989,998

**BUILDING AN INFORMATION MANAGEMENT SYSTEM FOR GLOBAL DATA SHARING: A STRATEGY FOR THE INTERNATIONAL LONG TERM ECOLOGICAL RESEARCH (ILTER) NETWORK**

**Kristin L. Vanderbilt<sup>1</sup>, David Blankman<sup>2</sup>, Xuebing Guo<sup>3</sup>, Honglin He<sup>3</sup>, Jianhui Li<sup>3</sup>, Chau-Chin Lin<sup>4</sup>, Sheng-Shan Lu<sup>4</sup>, Burke Chih-Jen Ko<sup>4</sup>, Akiko Ogawa<sup>5</sup>, Éamonn Ó Tuama<sup>6</sup>, Herbert Schentz<sup>7</sup>, Su Wen<sup>3</sup>, Bert van der Werf<sup>8</sup>**

<sup>1</sup>Sevilleta LTER, Albuquerque, New Mexico USA, <sup>2</sup>Israel LTER, Ben Gurion University, <sup>3</sup>Chinese Ecological Research Network (CERN), Chinese Academy of Sciences, Beijing, <sup>4</sup>Taiwan Ecological Research Network (TERN), Taiwan Forest Research Institute, Taipei, <sup>5</sup>Japan LTER, Kyoto, <sup>6</sup>GBIF Secretariat, Copenhagen, Denmark, <sup>7</sup>Umweltbundesamt GmbH, Vienna, Austria, <sup>8</sup>Alterra, Wageningen, Netherlands

**Abstract**

The International Long Term Ecological Research (ILTER) Network is a global network of sites arrayed in many ecosystems and countries that aims to address international ecological and socio-economic problems through collaborative research. To facilitate ILTER data discovery, access, and synthesis, a strategy for adopting common information management standards throughout the ILTER has been developed. The strategy proposes that ILTER use the Ecological Metadata Language (EML) standard in the short-term to establish a network-wide metadata catalog, while in the long-term ILTER should pursue the goal of an ontology-based information management system. This paper presents examples of the information management systems currently in use in the ILTER that are EML-based or based on a country-specific standard, and discusses possible mechanisms for accommodating the many different languages used throughout the ILTER. The advantages of ontology-driven information management systems are illustrated with examples from the ILTER, and the approach ILTER will take to realizing such a system for the whole network is discussed.

**Keywords:** LTER, ontology, metadata, semantic integration, language

**1. Introduction**

The International Long Term Ecological Research (ILTER) Network consists of 34 member countries that support long-term data gathering and analysis on a global scale to detect, interpret and understand environmental changes. The strategic plan for the ILTER Network of networks includes these ten-year goals:

1. Foster and promote collaboration and coordination among ecological researchers and research networks at local, regional and global scales
2. Improve comparability of long-term ecological data from sites around the world, and facilitate exchange and preservation of this data
3. Deliver scientific information to scientists, policymakers, and the public and develop best ecosystem management practices to meet the needs of decision-makers at multiple levels

To achieve these goals of collaboration, data compatibility, data exchange, and data preservation will require a significant investment in both ecological informatics research and

cyberinfrastructure development. Some ILTER networks have already invested in substantial technology infrastructure that uses different solutions for structuring, storing and analyzing data, and creating and managing metadata.

A viable ILTER information management solution would need to address these different system infrastructures to create an interoperable system of systems. It must also address the challenges of discovering and integrating data that are documented in different languages. To create such a system, ILTER information management and technology specialists from East-Asia Pacific, Europe, and North American regions have recommended that ILTER adopt Ecological Metadata Language (EML) in the short-term as the ILTER metadata standard in order to create a shared metadata catalog and data portal for the network. In parallel, the ILTER should engage in the ontology standardization process that will eventually support the semantic annotation of data.

In this paper, we review examples from the ILTER that illustrate how EML is already being successfully used and also how a country-specific metadata standard can be adapted to generate EML for inclusion in the ILTER metadata catalog. We also describe current examples of the implementation of ontologies within the ILTER, and outline the vision for ILTER's future ontology-driven information management system. We conclude with ILTER's strategy for engaging with the international standards development process to ensure interoperability within ILTER and between ILTER and other environmental networks such as the Global Biodiversity Information Facility (GBIF), whose fundamental operating principle is free and open access to biodiversity data.

## **2. Developing an ILTER Data Catalog: EML Implementation in the ILTER**

### **2.1 Why Choose EML as the ILTER Metadata Standard?**

EML is a standard for documenting ecological data that is implemented as a series of XML modules (EML Specification: <http://knb.ecoinformatics.org/software/eml/eml-2.0.1/index.html>). It has already been adopted by several ILTER networks (US LTER, Taiwan Ecological Research Network (TERN), Israel LTER, Mexico LTER, and South African Environmental Observation Network (SAEON)), because tools exist to create, manage, and analyze data using EML. The availability of these tools and the considerable experience of some ILTER personnel with this standard will make it easier for other ILTER members to adopt when their country or site initiates an information management system.

In order to create an ILTER-wide data catalog, all ILTER networks will generate "discovery-level" EML, a core set of elements including title, keywords, abstract, creator, and spatial and temporal domains. Each ILTER member network may choose to manage their metadata entirely as EML, or they may manage the bulk of their metadata in another system from which they generate discovery-level EML. Examples of both approaches are described below. ILTER will have to find solutions for handling metadata written in different languages in order to make all data accessible from a single portal, and mechanisms for dealing with this challenge are also discussed.

### **2.2 Examples of EML Usage and Generation in the ILTER**

*TERN EML-driven solution to carbon flux data management issues:* Currently, no universally accepted method of carbon flux data management has been established that uses a metadata approach for archiving, curating, discovering, retrieving and analyzing data. Instead, each flux research group has formed their own regional network such as CarboEurope, AmeriFlux, and AsiaFlux and each has developed software to address data management issues. Since 2004,

Taiwan Ecological Research Network (TERN) has collected existing EML-based tools and assembled them as a data management system that could be used universally in carbon flux research.

Using this EML-based data management system, a conceptual framework has been developed for flux data management that can be divided into three tiers (Figure 1). The first tier deals with datasets and related information. Data produced by eddy covariance sensors communicating automatically through wired or wireless networks are managed by this tier. In this first tier, all information related to a flux dataset is documented in EML using the Morpho EML editor (Higgins et al. 2002). The second tier relates to information management. Once metadata and data quality have been described and checked, the metadata are stored in the Metacat system (Java servlet, LDAP authentication, and backend schema-independent database). Data are stored using Storage Resource Broker (SRB) (Rajasekar et al. 2003), a data grid middleware software system. The third tier consists of web service based scientific workflows that allow easy access to the second tier. The Kepler scientific workflow system (Ludäscher et al. 2006) was used in this layer to model and execute the flow of data through a sequence of analytical steps.

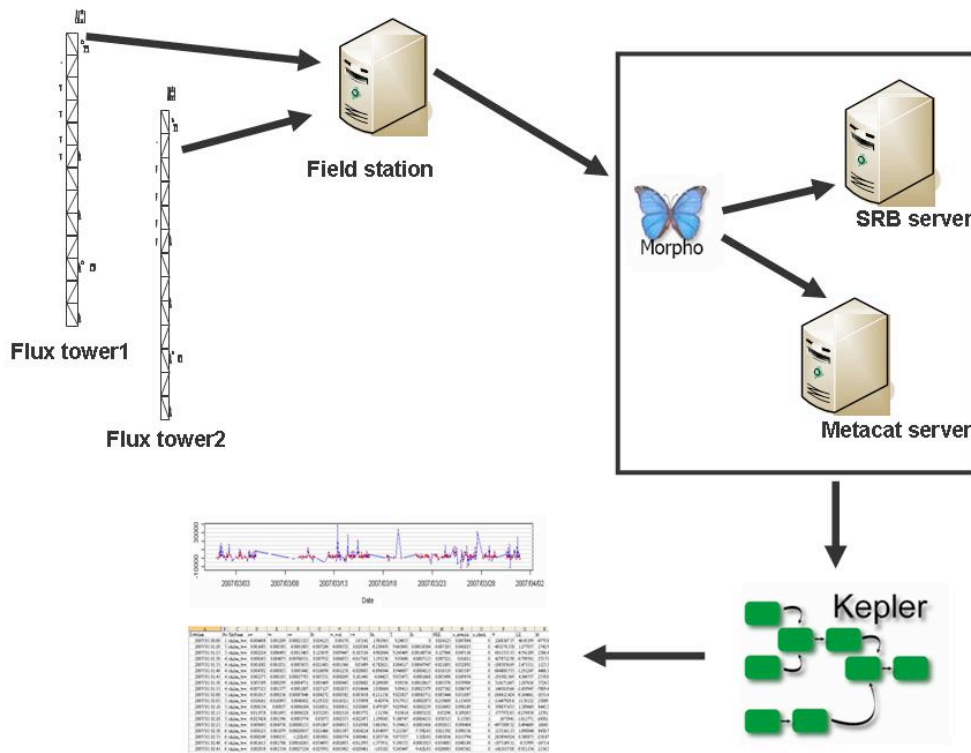


Figure1. Using EML-based tools for carbon flux data management

The use of this EML data management system was applied in Chilan, a TERN site where two flux towers have been set up since 2000. The two towers are equipped with vertical and horizontal wind vectors and the CO<sub>2</sub> mixing ratio at 20 Hz is measured with a sonic anemometer. A desktop computer collects these data. Every 30 minutes, the computer stores the raw data

which is downloaded weekly and loaded to a SRB server to be retrieved for analysis. Metadata for these raw data are created and stored in the Metacat. Then, using the Kepler system, five workflows are run that search metadata from the Metacat, download data from the SRB, rotate data coordinates, QA/QC the data, and create Web-Pearman-Leuning (WPL) corrections to standardize the flux data calculation process based on each 30 minutes of data collected. Output of the final calculation of all flux data are displayed in a text file which reports all the variables and a graphical file which shows the flux trend of a specific period. These secondary data can be saved locally or remotely.

The adaptation of the existing EML-based tools from the flux data management experiment has achieved the goal that sequential analyses of ecological data be accompanied by formal process metadata. Further, this data management system is portable, and can be used by any carbon flux research group.

*Adaptation of CERN metadata standard to generate EML:* In China, the Chinese Ecological Research Network (CERN) is the main organization conducting ecological research and data management, analysis, synthesis and sharing. Based on the draft of “The National Metadata Standard for Ecological Data Resources (GB/T 20533-2006)” that considered many metadata standards such as EML and ISO19115 while it was being developed, CERN proposed a metadata standard more relevant to CERN’s needs. This standard was adopted in 2006. Although the conceptual framework of CERN’s metadata has many elements, CERN trimmed many elements and only reserved those that were crucially necessary for describing ecological data when CERN built its physical metadata database.

CERN’s metadata database is composed of seven modules, including (1) dataset identification module, (2) entity identification module, which contains information about each entity (such as a database table or other file) in a dataset, (3) observational plot module, which describes each plot’s spatial coverage and geographic background information, as well as management information of the plot, (4) observational method module, (5) data quality evaluation module, (6) project information module, and (7) dataset distribution module. Although modules or elements in CERN’s metadata standard and EML are not exactly the same, a valid EML document can be generated from CERN’s system. Each EML element logically corresponds to one or more elements in CERN’s metadata. CERN’s identification module (Figure 2) includes elements that are quite similar to EML: dataset title, identifier, abstract, keywords, creator, date of dataset creation, status of the dataset, language, disk size of the dataset, spatial coverage, and temporal coverage.

The centralized CERN information management system harvests metadata in the CERN format from all CERN sites and stores it in an Oracle RDBMS and provides a central metadata catalog for all CERN data. Metadata content can be output to XML documents, and CERN can generate EML documents to be included in the ILTER metadata catalog.



♀ 数据集标识符 dataset identifier
♀ 数据集名称 dataset name
摘要 dataset abstract
目的 dataset purpose
创建者 creator
其它贡献者 contact person
发布日期 dataset publishing date
状态 status of dataset
语种 language
字符集 charset
存储量 disk size of dataset
记录数 number of records
关键词 keyword set
开始日期 beginning date of the temporal coverage
结束日期 ending date of the temporal coverage
地理边界矩形之西部边界经度 longitude of west boundary of the spatial coverage
地理边界矩形之东部边界经度 longitude of east boundary of the spatial coverage
地理边界矩形之北部边界纬度 latitude of north boundary of the spatial coverage

Figure 2. CERN metadata's identification module contains elements found in EML discovery level metadata.

### 2.3 ILTER Confronts Metadata Language Issues:

EML harvested from different sources may be documented in different languages and character systems, and development of an ILTER-wide data catalog will require that all metadata be represented in one language (assumed to be English for purposes of this paper). Three mechanisms for addressing the language issue are discussed below: 1) Internationalization of the metadata exchange format, 2) Localization of the software tools used, and 3) Creation of a multilingual thesaurus.

*Internationalization of the metadata exchange format:* One option for resolving language issues is to generate multiple EML documents, one in English and one in the ILTER Network member's native language. This approach requires the most time and work, but has the advantage of maximizing the semantic integrity of each EML document. A second option would be to include multiple languages in a single EML element. Japan LTER, for instance, would include Japanese and English titles in the <title> element. TERN currently puts both English and Chinese into the <title> (Figure 3) and <abstract> elements of EML.

<title> **Using Genetic Algorithm to Predict Distribution  
of Taiwan Fir** 運用基因演算法預測台灣冷杉分佈 </title>

Figure 3. Internationalization of the <title> EML element, showing the title in English and Chinese.

A third, but least satisfactory option, would be to include duplicate elements for discovery-level EML, one for English and one for the native language. The drawback to this approach is that the title element will no longer be unique, and could confuse the query engine.

*Localization of the software tools:* Scientists will want to document their data and interact with metadata tools in their native language. The software design should separate the presentation layer from the logic layer to provide this localization capability. By providing software skins in different languages, people from different nations will be comfortable using the software. For instance, the Metacat metadata database was developed with an English language user interface, but has been localized to a Chinese version by scientists at TERN (Lin et al. 2008) (<http://metacat.tfri.gov.tw/tfri/>). The 1.7.0 version of Metacat was altered so that the display language is Unicode which supports two-byte script systems such as Chinese. The updated version (Metacat 1.8.1) has been defaulted to be able to display all language codes, including Chinese.

*Development of a multilingual thesaurus:* Each scientific domain should standardize a controlled vocabulary which can then be the basis for a domain thesaurus which maps semantically equivalent terms. Equivalent terms could be translated between languages, as has been done for the General Multilingual Environmental Thesaurus (GEMET) in Europe (<http://www.eionet.europa.eu/gemet>).

### **3. Advanced Data Integration Using Ontology-Driven Systems**

#### **3.1 Need for Ontology Systems**

The metadata catalog described thus far will be effective to broadcast the availability of ILTER data holdings, but it does not fully solve goals (2) and (3) of the ILTER strategic plan. To achieve data compatibility across ILTER systems will require a full understanding of the semantics of data holdings in each ILTER program. Thus, ILTER will have a strong need to participate in ontology development. An ontology is a description of a set of concepts and the relationships between them that enable data discovery and integration. The ontology system will help ILTER achieve its 10 year goals by supporting global data syntheses through the development of semantic data discovery services, semantic data integration services, and data access services that leverage data semantics.

Although a complete framework for semantic data integration does not yet exist within free software available for ecologists, we present two examples below that show how this concept can be realized and the additional power to find and integrate data that ontologies offer. The first example is the MORIS system developed by LTER Europe and the second is SeMIS, developed at the Chinese Academy of Sciences (CAS). We then describe a vision for the ontology system for the whole ILTER.

#### **3.2 Example Ontology System Successes in the ILTER**

*MORIS:* The MORIS system (<http://www.umweltbundesamt.at/en/umweltschutz/oekosystem/informationssystem/>) demonstrates the successful use of ontologies within a single framework (Schentz and Mirtl 2003). Version 1 of MORIS is an information system primarily designed for the Austrian part of the United Nations Economic Commission for Europe (UN – ECE) project “Integrated Monitoring”, dealing with extremely heterogeneous observations on soil, vegetation, water, and air. In MORIS, metadata are part of the ontology, describing the

meaning of observations and measurements in detail with scientific concepts and relations between them. Types of scientific concepts include observation design, parameters, observed entities, methods, treatments, samples, context of observation, people, institutes, and projects. Those ontologies are closely coupled with the measured data values so that scientists accessing the data for synthesis and analysis can access data and metadata through a uniform interface and correctly interpret them.

One of the main differences between the MORIS system and EML is the treatment of methods. In EML, methods are described in natural language. In MORIS each part of a method is described using a controlled vocabulary and a system of relationships. Because methods are defined using an ontology, the MORIS system can determine whether or not two sets of research data can be integrated without requiring the researcher to compare two text documents.

*SeMIS: A semantic-based metadata integration system for scientific data:* In contrast to the fully-integrated ontology system used by LTER Europe, the Semantic-based Metadata Integration System (SeMIS) demonstrates the successful use of ontologies for integrating data from heterogeneous existing metadata systems. SeMIS, developed by researchers at the Chinese Academy of Sciences (CAS), is a framework that enables the translation of metadata formats that conform to different standards to a global schema so that multiple metadata standards can be accessed, queried and manipulated in an integrated way. Using a domain ontology, metadata can be manipulated in a uniform way based on the common semantics of metadata from different standards regardless of the differences in metadata format and structure.

SeMIS was developed in three steps. First, a global domain ontology was developed by domain experts and computer experts working together. The global ontology has two roles: 1) It provides the user access to the data with a uniform query interface to facilitate the formulation of a query on all the metadata sources, and 2) It serves as the mediation mechanism for accessing the distributed data through any of the metadata sources. Second, metadata elements were mapped to the concepts in the global ontology. Metadata can be originally encoded and expressed in XML format or stored in relational database or some data grid system such as SRB. For the XML format, the path-to-path mapping strategy was used where XPath was mapped to ontology classes and/or property paths. The generated mapping rules were stored in a mapping table. Finally, based on the mapping table built in the previous step, the actions of manipulating the ontology are translated to the actions of manipulating metadata. For example, semantic queries are rewritten into XQuery on each local XML dataset, and then the returned query results are reformatted for end users.

SeMIS is useful for integrated metadata browsing and searching and has been taken into practical use in the Qinghai Lake CERN research site investigation and research database project. Based on an observation ontology, users can easily browse and search animals and plants living in specific environments. Currently, SeMIS mainly integrates XML encoded metadata and researchers are working to make SeMIS support more metadata formats.

### **3.3 A Vision for an ILTER Ontology System**

These examples from LTER Europe and CAS illustrate the advances that can be made through ontological modeling of environmental data. However, for all of ILTER to take full advantage of such a system, ILTER would need to engage in ontology standardization efforts that are occurring within the broader ecological informatics community and build interoperable semantic service implementations across the whole network.

Madin et al. (2008) characterize ontologies as framework ontologies and domain ontologies. Domain ontologies provide the detailed semantic information associated with a particular discipline, such as in sub-disciplines of ecology. For example, Williams et al. (2006) created a domain ontology to describe the specifics of food-web interactions among species. However, if domain ontologies are developed in isolation, they may be difficult or impossible to integrate due to logical inconsistencies in their modeling approaches. Thus, framework ontologies that provide a common modeling perspective and that can be used to integrate extended domain ontologies are critical. One such framework ontology is the Extensible Observation Ontology (OBOE) (Madin et al. 2007). OBOE provides a common modeling framework that can be used to create specialized domain ontologies that address specific aspects of scientific observations, such as what entity was measured in an observation, the characteristic of that entity that was measured, and the context in which the measurement occurred. Another framework ontology is ALTERNet Core ([http://www5.umweltbundesamt.at/ALTERNet/index.php?title=Ont:ALTER-Net\\_Ontology](http://www5.umweltbundesamt.at/ALTERNet/index.php?title=Ont:ALTER-Net_Ontology)), which models the observation in a similar way. The ILTER needs to participate in the development of one comprehensive ontology framework for observational science data.

Once an ontology framework and a set of domain ontologies are available, ontology terms can be associated with data collected in the field by mapping ontology fields onto data measurements, a process termed 'semantic annotation' (Madin et al 2008, Bowers and Ludäscher 2003, Bowers et al. 2004). Such semantic annotations allow software systems to use an ontology for data discovery and integration and then access the associated data via the annotation (e.g., use of semantics in workflow design (Berkley et al. 2005)).

Even with a global ontology framework and broadly accepted domain ontologies, we expect ILTER sites will need to maintain their existing local infrastructures because of the significant investment they represent. Thus, the software architecture for an ILTER ontology system must accommodate those systems by allowing the global ontology to be connected to those local systems. One possible architecture would make use of a mediator whose function is to query local systems based on a mapping between the local (ontology) and the global ontology. In such an integrated system, the mediator process is a query/integration engine that exposes the local data via the concepts in the global ontology. The advantage of this integrated architecture is that the local data-infrastructure need not be changed. Only the mediator between the local infrastructure and ontology needs to be created. We expect that when more and more people use the ontology, the local data infrastructure gradually will adopt and adapt concepts from the global ontology. This will also lead to standardization and unification of concepts in the ILTER community.

#### **4. An International Community for Developing Ontology Standards**

Data integration within the ILTER network and between ILTER and other biodiversity networks can be enhanced by the adoption of common framework and domain ontologies. These include, e.g., the aforementioned OBOE and CEDEX ([http://www.umweltbundesamt.at/umweltdaten/schnittstellen/cedex/cedex\\_protege/?&templ=1](http://www.umweltbundesamt.at/umweltdaten/schnittstellen/cedex/cedex_protege/?&templ=1)) both of which provide framework ontologies for ecological data. Biodiversity Information Standards (BIS) TDWG (<http://www.tdwg.org>), a primary, international organization dealing with standards for exchange of biodiversity data, is now also focusing on ontologies. It has adopted a new technical architecture (<http://wiki.tdwg.org/TAG/>) with ontologies and globally unique identifiers for biodiversity objects based on Life Science Identifiers (LSIDs) as core components. The move to

ontologies was adopted to overcome the limitations of defining standards through XSD schemas which are document-centric, difficult to extend, and difficult to integrate across schemas. By adopting an object-based / ontology approach, common concepts can be defined and reused across different communities and still be expressed in community-specific XSD schemas if required.

TDWG has begun the process of defining LSID vocabularies (<http://wiki.tdwg.org/twiki/bin/view/TAG/LsidVocs>) by expressing common biodiversity concepts as found in various XSD based schemas (e.g., Darwin Core, Taxon Concept Schema, Natural Collections Descriptions) in a formal ontology language (OWL). There is an urgent requirement to provide additional vocabularies for scientific observations – a domain of particular relevance to ILTER, and this is currently being undertaken via the formal TDWG standards process by an Observational Data Task Group which is developing a core semantic model for observational data in the ecological and environmental sciences. Through active participation by both scientists with domain knowledge and technical personnel, the ILTER community can contribute to, and benefit from, these efforts. The outcome will be enriched LSID vocabularies for observations that build on and extend the set of TDWG LSID vocabularies, and can be deployed in the ILTER information management system to enable enhanced data discovery, interpretation and integration both within and across disciplines.

## 5. Summary

The goal of the ILTER information management system is to foster broad-scale research synthesis efforts by facilitating the discovery, access and integration of global data resources. In the short-term, ILTER will establish a data catalog based on EML contributed by all ILTER member countries. To achieve the long-term vision of an ontology-driven ILTER information management system, ILTER scientists will also participate in the development of the semantics necessary for the creation of standard framework and ecological and socio-ecological domain ontologies that will be used to support data integration within the ILTER and between the ILTER and other organizations.

## Acknowledgements

This paper is a product of the "ILTER Information Management Workshop on Ontology/EML Integration" that was held at Lake Taihu Field Station, China, April 7-12, 2008. This workshop was supported by the Chinese Ecological Research Network (CERN) . The US National Science Foundation supported travel by US participants to attend this workshop. We thank two anonymous reviewers for comments that improved this paper.

## References

- Berkley, C., Jones, M.B., Bojilova, J., and Higgins, D., 2001. Metacat: a schema-independent XML database system. Proc. of the 13<sup>th</sup> Intl. Conf. on Scientific and Statistical Database Management. IEEE Computer Society.
- Fegraus, E.H., Andelman, S., Jones, M.B., and Schildhauer, M. Maximizing the value of ecological data with structured metadata: an introduction to ecological metadata language (EML) and principles for metadata creation. Bulletin of the Ecological Society of America. July 2005: 158-168.
- Higgins, D., Berkley, C., and Jones, M.B., 2002. Managing heterogeneous ecological data using Morpho. Proc. of the 14th Intl. Conf. on Scientific and Statistical Database Management.

- Lin, C.C., Porter, J.H., Lu, S.S., Jeng, M.R., and Hsiao, C.W., 2008. Using structured metadata to manage forestry research information: a new approach. *Taiwan J. For. Sci.* 23: 133-43.
- Ludäscher, B., Altintas, I., Berkley, C., Higgins, H., Jaeger, E., Jones, M., Lee, E.A., Tao, J., and Zhao, Y., 2006. Scientific workflow management and the Kepler system. *Concurrency Comput. Pract. Exp.* 18 (10): 1039–1065.
- Madin, J.S., Bowers, S., Schildhauer, M.P., and Jones, M.B., 2007. Advancing ecological research with ontologies. *Trends in Ecology and Evolution* 23:159-168.
- Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., and Villa, F., 2007. An ontology for describing and synthesizing ecological observation data. *Ecological Informatics* 2: 279-296.
- Rajasekar, A., Wan, M., Moore, R., Schroeder, W., Kremenek, G., Jagatheesan, A., Cowart, C., Zhu, B., Chen, S.Y., and Olschanowsky, R., 2003. Storage Resource Broker - Managing Distributed Data in a Grid. *Computer Society of India Journal, Special Issue on SAN* 33(4): 42-54.
- Schentz, H., and Mirtl, M., 2003. MORIS: a universal information system for environmental monitoring. In: Schimack, G.P. (Ed.), *Environment Software Systems*, vol. 5. Springer.
- Williams, R.J., Martinez, N.D., Golbeck, J., 2006. Ontologies for ecoinformatics. *J. Web Semant.* 4: 237–242.

## VEGA: A FLEXIBLE DATA MODEL FOR ENVIRONMENTAL TIME SERIES DATA

L. A. Winslow<sup>1</sup>, B. J. Benson<sup>1</sup>, K. E. Chiu<sup>3</sup>, P. C. Hanson<sup>1</sup>, T. K. Kratz<sup>2</sup>

<sup>1</sup>Center for Limnology, University of Wisconsin-Madison, 680 N. Park Street, Madison, WI 53706 USA; <sup>2</sup>Trout Lake Station, Center for Limnology, University of Wisconsin-Madison, 10810 County Highway N, Boulder Junction, WI 54568 USA; <sup>3</sup>State University of New York at Binghamton, P.O. Box 6000, Binghamton, NY 13902

### Abstract

As large scale sensor networks grow, effective data curation of large data volumes is becoming important. Many sites have filled this need with site-specific database systems and software. Within the Global Lake Ecological Observatory Network (GLEON), a fundamental need was identified for a data model that allows for growth and flexibility in sensing platforms and configurations requiring minimal or no data model changes. The Vega data model, a variant of the Observation Data Model developed by CUASHI, is designed to fulfill that need. The Vega data model is a flexible, site agnostic data model optimized for high temporal resolution ecological sensor network data sampled at frequencies as high as a few seconds. Instead of storing data in a spreadsheet-like format with different variables denoted by columns, Vega stores observed values individually and describes them fully with linked, metadata-containing tables. While being difficult to intuitively recognize, this more flexible and portable data model is beneficial at the individual institution level because it handles additional sensor deployments and configuration changes with no change in database structure and at an inter-institution level because it represents a portable standard against which flexible and site-agnostic software can be developed. Deployment and testing of this system has already begun within GLEON and has involved five different institutions.

### 1. Introduction

Modern ecological sensor networks are growing at a rapid rate (Porter et al. 2005). Several large scale projects such as NEON (National Ecological Observatory Network), WATERS (Water and Environmental Research Systems Network), and OOI (Ocean Observatories Initiative) propose deploying large scale environmental sensor networks across the US. Up to now, most groups with data curation systems have implemented site specific custom structures. While there are many different structures that effectively curate sensor network data, it is challenging to balance ease of use, query performance, and flexibility. Some groups have attempted to address the flexibility challenge by creating structures that store observations individually. One prominent example of an observation-based structure is the Observation Data Model (ODM) designed by CUAHSI (Consortium of Universities for Advancement of Hydrologic Science) (Horsburgh et al. 2008). In this paper, we describe a variant of the ODM called Vega. Vega was inspired by the ODM and has borrowed from ODM's terminology and concepts. The Vega data model is an observation-based data model for high-resolution time series data sampled at frequencies as high as a few seconds and is designed to optimize performance, flexibility, and simplicity.

It is beyond the scope of this paper to describe in detail the differences between Vega and ODM. However, the major differences are the separation of observation metadata into the

‘Streams’ table, additional indexing for performance, database level mechanism for enforcing data uniqueness, and simplified metadata table structure.

Vega is currently being implemented in the Global Lake Ecological Observatory Network (GLEON; [gleon.org](http://gleon.org), Kratz et al. 2006). GLEON is an international, grassroots network of limnologists, ecologists, engineers, and information technology experts who have a common goal of building a scalable, persistent global network of lake ecology observatories. Data from these observatories will allow us to better understand key processes such as the effects of climate and land-use change on lake function, including carbon cycling in lakes, and the role of episodic events, such as major rainstorms and hurricanes/typhoons, in resetting lake dynamics. The current deployed observatories consist of instrumented platforms on lakes around the world that are capable of sensing key limnological variables and moving the data in near-real time to web-accessible databases. Vega was developed after recognizing a fundamental need of GLEON for a data model that is flexible in both the number and configuration of instrumented sites and portable between institutions. Vega has been designed to meet the goals of GLEON but is applicable to any system with time series data.

### 1.1 Data Model Structure

The Vega data model stores data as individual observed values. All values are stored in a single large table and are directly linked to their supporting metadata. Through careful use of normalization and indices, redundancy is reduced to save space and query performance is optimized to improve retrieval times. Below is a description of the table, relationship, and index structure as well as conceptual issues in the representation of the data.

### 1.2 The Value

The fundamental record in Vega represents individual floating point values of discrete measurements. These values are all stored in a single table called ‘Values’. Because this table stores all observations within the system, it must be optimized to minimize storage space requirements and enhance query performance against it.

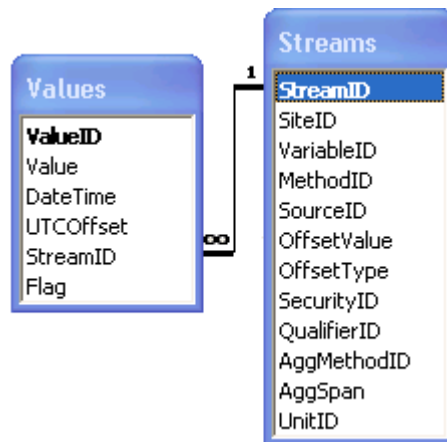


Figure 7 - ‘Values’ and ‘Streams’ tables at the core of Vega with the one-to-many relationship between them.

Each value is stored as a floating point double, is time stamped by a date-time field, and is linked to its metadata by its stream identification (Figure 7), to be described in more detail



later. The 'ValueID' field is included for convenience when programming and manipulating individual values. 'Flag' is included to allow for QA/QC descriptive data and to maintain backwards compatibility with systems that use data flagging as an indicator of potential data quality or other metadata.

Duplicate data are prevented at the table level. A unique index is defined for the 'Values' table on the 'DateTime' and 'StreamID' columns. No two values can have both the same stream and timestamp.

### 1.3 The Stream

The data stream is an entity designed to fully describe data that only vary through time, or in other words, a unique time series. Each stream is described by attributes stored in the 'Streams' table and can be thought of as a unique combination of attributes. For example, air temperature sampled at a particular meteorological station through time would be a unique stream. Soil temperature at that same station would be a different data stream. Unique time series coming from multivariate sensors must be separated and described individually as different variables or as the same variable differentiated on a different dimension. For example, thermistor chain data recorded at multiple depths would return multiple unique time series each described as water temperature but differentiated by their depth.

Each stream has required attributes and optional attributes necessary when those required are insufficient to uniquely describe the stream or when additional metadata are desired. Each stream is assigned a unique integer identifier, 'StreamID', forming the one-to-many relationship back to the 'Values' table. Duplication of streams is prevented by a unique index spanning all columns except 'StreamID'. Most of these attributes are stored as foreign keys, linked to other supporting tables and not directly in the 'Streams' table with only 'OffsetValue' and 'AggSpan' being exceptions. In the future, the entity 'Stream' may be renamed to avoid ambiguity, especially in aquatic contexts.

### 1.4 Uniqueness

Inadvertent insertion of duplicate data is prevented by table-level unique indexing applied to the 'Streams' and 'Values' tables. Programs inserting data do not need to know what data have been entered into the database as all potential duplicate inserts are prevented.

### 1.5 Supporting Attributes and Entities

The stream attributes map onto logical and useful supporting entities. Most of these are relatively straightforward, like the concepts of *site* and *variable*, but some are more abstract, like *aggregation* and *method*. The definitions are restricted to maintain usefulness and reduce ambiguity but are kept flexible to support different uses and unique requirements. For example, defining site too rigidly as specific latitude/longitude coordinates could make simple queries unwieldy.

Data aggregation is described by the *aggregation method* and *aggregation span* fields. Aggregation method is the method by which the data are aggregated, e.g., the mean, maximum, or standard deviation of a signal over a certain time. Aggregation span is the timeframe over which data are aggregated, e.g., one hour, one day. This allows simple types of temporal aggregation to be represented in the database but does not attempt to include other potential aggregation types (e.g., spatial). If the data are sampled instantaneously and not aggregated, the

aggregation method is defined as instantaneous and the span is zero. Aggregation types are stored in the 'AggTypes' table.

Sites in Vega are generally 2d locations in space. For some purposes, a looser definition of site may be useful. For example, in limnology, it may be helpful to define a whole lake as one site. Sites are stored in the 'Sites' table and can have a name, latitude and longitude, elevation, and country. Only the name is required. A third dimension is available through the 'OffsetType' and 'OffsetValue' fields.

An offset can be used to describe many different situations where a simple 2d site doesn't adequately or uniquely describe a value. Offset value can be any double floating point value and offset type can be of the user's choosing. Common examples include depth, height, and distance along a transect. Offset types are stored as name/id pairs in the 'OffsetTypes' table.

The 'VariableID' is linked to the 'Variables' table and describes what type of measurement or observation the value describes. The 'Variables' table stores name/id pairs.

'Method' within Vega is used to associate each stream with either a unique sensor or laboratory method attribute. This stream attribute is optional and is not required to fully describe the data, but can be useful for sites with a large number of sensors and complex calibration requirements.

Each value's unit is included and is linked to the 'Units' table by the 'UnitID' field. Units are included in defining stream uniqueness as it is possible that the same variable is stored with differing units.

## 2. Discussion

The Vega data model is a flexible system for storing environmental time series data and can handle changes to deployments and configuration without structural change or database level manual intervention. It is currently in use at five GLEON member sites and centrally as an intersite repository of shared data. Many interesting implications of using an observation level data model and Vega specifically have arisen over this time.

The Vega data model has advantages over more traditional archival models. Individual sensor deployments can be added and removed easily and without requiring manual intervention in the database system. We have used this system to store data from not only long-term sensor deployments, but also short-term datasets generated by individual experiments. Tools developed for Vega will always work against the same set of tables, regardless of what data are contained. Tools need not be changed or updated when datasets change. Vega also offers flexibility in discovering data. For example, it is very straightforward to select a site and determine variables measured at that site, or select a variable and determine the sites where that variable has been measured. Potential exists for even more pathways to discovering data, like retrieving all data sampled by a single sensor that was deployed to multiple different sites to get sensor history and quality assurance statistics.

Vega's flexibility not only proves to be an important advantage in single institution systems but also presents a great opportunity within the entire community. When flexible standards are adopted, all systems and software developed for those standards can be adopted by other institutions, reducing the burden of site-specific software development. Just as any browser that understands the HTML standard is immediately compatible with sites developed in HTML, using a common data storage standard like Vega across institutions would provide portability to tools that are developed as well as compatibility among systems that use the standard.

The Vega data model also has some limitations. Each value is individually time stamped and indexed. This means that observations don't share timestamps or unique indexes as is typical for a flat table structure and thus more storage space is required comparatively. Data are not inherently in the form most users are used to seeing so tools must be developed to expose data stored in Vega in a form readily consumed by the user. Tools can, depending on the circumstances, be more difficult to develop. The data model is inherently more complex than a simple flat table view of the data as it contains a number of table relationships and more abstract table definitions. This structure may not be intuitively recognizable to the average user and requires further documentation and explanation.

Additionally, because the values are individually time stamped, data collected at the same time from the same platform must be transformed to matrix format if one wants a spreadsheet-like view of multiple variables. This pivoting is currently done programmatically by the query tools based on the user's demands. The separation of streams and values makes editing data somewhat more complicated. Tools for editing values or altering metadata may be required to alter individual values, move values from one stream to another, or alter stream information independently. For example, changing the depth for a series of values may create a conflicting stream, in which case the 'StreamID' for values would need to be altered instead of simply changing the offset value.

Vega has not yet undergone rigorous experimental testing but is the subject of an ongoing case study through its use in GLEON. During this time a few performance characteristics of interest have come to light. Simple data retrieval query times depend more on the number of values retrieved and less on the number of values stored in the 'Values' table. Well-formed queries retrieving less than 100,000 values typically execute in less than one second. Inserting data typically takes, per value, longer than retrieval but still approaches tens of thousands of values per minute, very reasonable for most purposes. Storage requirements have been very reasonable. While naturally requiring more space than storage in raw text files, Vega has reliably required only about 100MB per million values stored, which represents about 19 variables sampled every ten minutes for one year. These metrics were measured on a commodity, dual-core based server and are provided as anecdotal evidence only.

As discussed, because of the more complex nature of Vega's data structure, it is important that tools are developed to aid users in data importation, editing, and querying. Several tools have already been developed and are being improved to fulfill this role. For discovering and retrieving stored data, an ASP.net web application called dbBadger (<http://dbbadger.gleonrcn.org>) was developed. For parsing and importation of data, GLEONDN (<http://gleon.org/index.php?pr=Products>) was developed. Both of these applications have been in use in one form or another for several months, are open source, and are freely available. Additional tools are either being planned or are currently in development. These tools include dynamic figure and data output systems for use in public web sites, data management and editing tools for manual QA/QC, and programmatic query tools for dynamic access to data by models and statistical tools.

Up to this point, Vega has been developed on a MySQL 5.0 backend. The data model in SQL form can be downloaded from the GLEON website (<http://gleon.org/index.php?pr=Products>). In the future, the Vega development group hopes to expand and interact with a broad range of groups inside and outside GLEON. The data model has already benefited from the expertise and input of GLEON network. This input has driven Vega, and especially the tools

built around it, in the direction they are headed today. As an evolving technology, Vega will change and be updated to meet new requirements and match emerging standards.

### **3. Conclusion**

By storing observed values individually and describing individual time series as streams, we created a flexible data model that fulfills the needs of GLEON and potentially other groups with similar curation requirements. Despite being somewhat less intuitive and having increased software development overhead, Vega's advantages in flexibility and portability make it a competitive alternative to traditional site-specific data curation systems.

### **Acknowledgments**

We thank Tony Fountain for his suggestion to consider ODM as an appropriate data model for GLEON. We also thank the developers of the ODM, without which Vega would not have been developed. The National Science Foundation (grants DEB-0217533, DBI-0639229, and DBI-0446017) and the Gordon and Betty Moore Foundation supported this work. We thank the member GLEON sites for their cooperation and encouragement.

### **References**

- Horsburgh, Jeffery S., D. G. Tarboton, David R. Maidment, and Ilya Zaslavsky. 2008. A relational model for environmental and water resources data. *Water Resources Research* 44, W05406, doi:10.1029/2007WR006392.
- Kratz, Timothy K., Peter Arzberger, Barbara J. Benson, Chih-Yu Chiu, Kenneth Chiu, Longjiang Ding, Tony Fountain, David Hamilton, Paul C. Hanson, Yu Hen Hu, Fang-Pang Lin, Donald F. McMullen, Sameer Tilak, Chin Wu. 2006. Towards a Global Lake Ecological Observatory Network. *Publications of the Karelian Institute* 145:51-63.
- Porter, J., P. Arzberger, H. Braun, P. Bryant, S. Gage, T. Hansen, P. Hanson, F. Lin, C. Lin, T. K. Kratz, W. Michener, S. Shapiro, and T. Williams. 2005. Wireless sensor networks for ecology. *Bioscience* 55:561-572.

## **LEEASP: A LINKED ENVIRONMENT OF COORDINATED MULTIPLE VIEWS FOR EXPLORATORY ANALYSIS OF LARGE-SCALE SPECIES DISTRIBUTION DATA**

**Jianting Zhang<sup>1</sup>, Kate S. He<sup>2</sup> and Michael Gertz<sup>1</sup>**

<sup>1</sup> Department of Computer Science, University of California at Davis, Davis, CA 95616

<sup>2</sup> Department of Biology, Murray State University, Murray, KY 42071

### **Abstract**

Exploratory analysis of large-scale species distribution data is essential to gain information and knowledge, stimulating hypotheses and seeking possible explanations of species distribution patterns. In this study, we report our design and implementation of LEEASP, a Linked Environment for Exploratory Analysis of large-scale Species Distribution data. LEEASP utilizes state-of-the-art advanced visualization techniques and multiple view coordination techniques to visualize different data sources that are relevant to species distribution data analysis and interact with users in a coordinated manner. As a case study, range maps of tree species in North America compiled by Elbert Little are used to derive the geographical and taxonomic data in the experiments. Environmental data derived from WORLDCLIM datasets and the EPA Level III Ecoregion data for the same study area are used as well to demonstrate the capabilities of the prototype system.

**Keywords:** Species Distribution, Exploratory Analysis, Large-Scale, Coordinated Multiple Views, GIS, Graph Visualization

### **1. Introduction**

Quantifying species-environment relationships, i.e., how plants and animals are distributed on the Earth in space and time, has been one of the important questions studied by biogeographers and ecologists. The availability of species distribution and the associated environmental data has increased significantly in recent years due to technological advances. While most of the current studies on species distribution modeling and prediction focus on a single species or a small number of selected species, the capabilities of exploratory analysis of large-scale species distribution data are essential to gain information and knowledge, stimulating hypotheses and seeking possible explanations of species distribution patterns.

Tree and graph visualization techniques are very useful in exploring species taxonomies (e.g., Parr et al. 2007). However, the taxonomic data in the studies are not linked to the geographical distributions and no geospatial exploration is involved. On the other hand, most existing GIS-based species distribution mapping techniques treat the distribution of an individual species as a separate layer. However, layers in most GIS systems (e.g., ArcGIS) are flat and the linear structure (list) cannot reflect the hierarchical taxonomic relationships among taxa. Subsequently, results of spatial or attribute-based queries are displayed as a stack of layers where the taxonomic relationships of species represented by the layers are difficult to explore. It is generally very cumbersome to manipulate layers at the order of hundreds even in a powerful commercial GIS and virtually impossible to use the approach when exploring the distributions of tens of thousands of species if they are treated as individual layers in a GIS.

In this paper, we report the design and implementation of LEEASP which utilizes the state-of-the-art advanced visualization techniques and multiple view coordination techniques to

visualize different data sources that are relevant to species distribution data analysis. LEEASP takes rasterized species distribution data for each species and environmental records associated with the raster cells. It computes the species distributed in each cell and arrange them into a tree form (termed taxonomic tree). The taxonomic tree over the whole geographical region under study (or study area) is the union of the sub-trees of all the cells in the region and consist all the species distributed in the area. The cells in the geographical domain are thus linked to the trees in the taxonomic domain. Compared to the linear list structure, the tree structure (a special form of a graph structure) is more suitable for visualizing and analysis of large and complex application-specific data that can be linked to geographical data. LEEASP integrates the tree/graph-based visualization techniques and GIS functionality for exploratory analysis of large-scale species distribution data. The work is the continuation of our previous work on extending open source Java GIS for exploring ecoregion-based biodiversity data (Zhang et al. 2007).

In this study, the range maps of tree species in North America compiled by Elbert Little and digitally available at USGS website (<http://esp.cr.usgs.gov/data/atlas/little/>) are used to derive the geographical and taxonomic data. Environmental data are derived from WORLDCLIM (Hijmans et al. 2005) and the EPA Level III Ecoregion data ([http://www.epa.gov/wed/pages/ecoregions/level\\_iii.htm](http://www.epa.gov/wed/pages/ecoregions/level_iii.htm)) are used as well to demonstrate the capabilities of LEEASP.

## 2 The Prototype: Design and Implementation

Four data views, namely the geographical data view, the environmental data view, the taxonomic data view and the ecoregion data view, have been implemented in LEEASP to visualize geographical, environmental, taxonomic and ecoregion data. They are referred to as the *Geographical* view, the *Environmental* view, the *Taxonomic* view and the *Ecoregion* view for short, respectively. A screen snapshot of the prototype is shown in Fig. 1. The *Geographical* view visualizes the geographical distributions of one or multiple species by embedding a GIS. The *Environmental* view displays the environmental envelopes (minimum and maximum values) of a subset of cells and the values of the environmental variables of the cells. The *Taxonomic* view displays all the species distributed in a region and their higher taxa in a tree format. The sub-trees representing species distributed in a subset of the cells can be highlighted. Finally the *Ecoregion* view displays the hierarchy of the ecoregions in the study area in a tree format and the paths from the root to the ecoregions containing the subset of selected cells can be identified.

Besides the basic display or visualization functions in the four views, an important feature of the prototype system is that the four views are linked and coordinated to provide richer functionality. In Fig. 1, when the tree node **F=Platanaceae** (F stands for Family) is selected in the *Taxonomic* view, the cells associated with the node are selected in the *Geographical* view and highlighted (in yellow) to help users understand the distributions of the taxa. The paths of the tree in the *Ecoregion* view that identify all of the most detailed ecoregions (Level III) by which the cells fall in geographically, are color-coded (in red) to help users explore the relationship between taxa and ecoregions. Finally the *Control* and the *Summary* panels in the *Environmental* view are populated with the environmental envelopes of the cells.

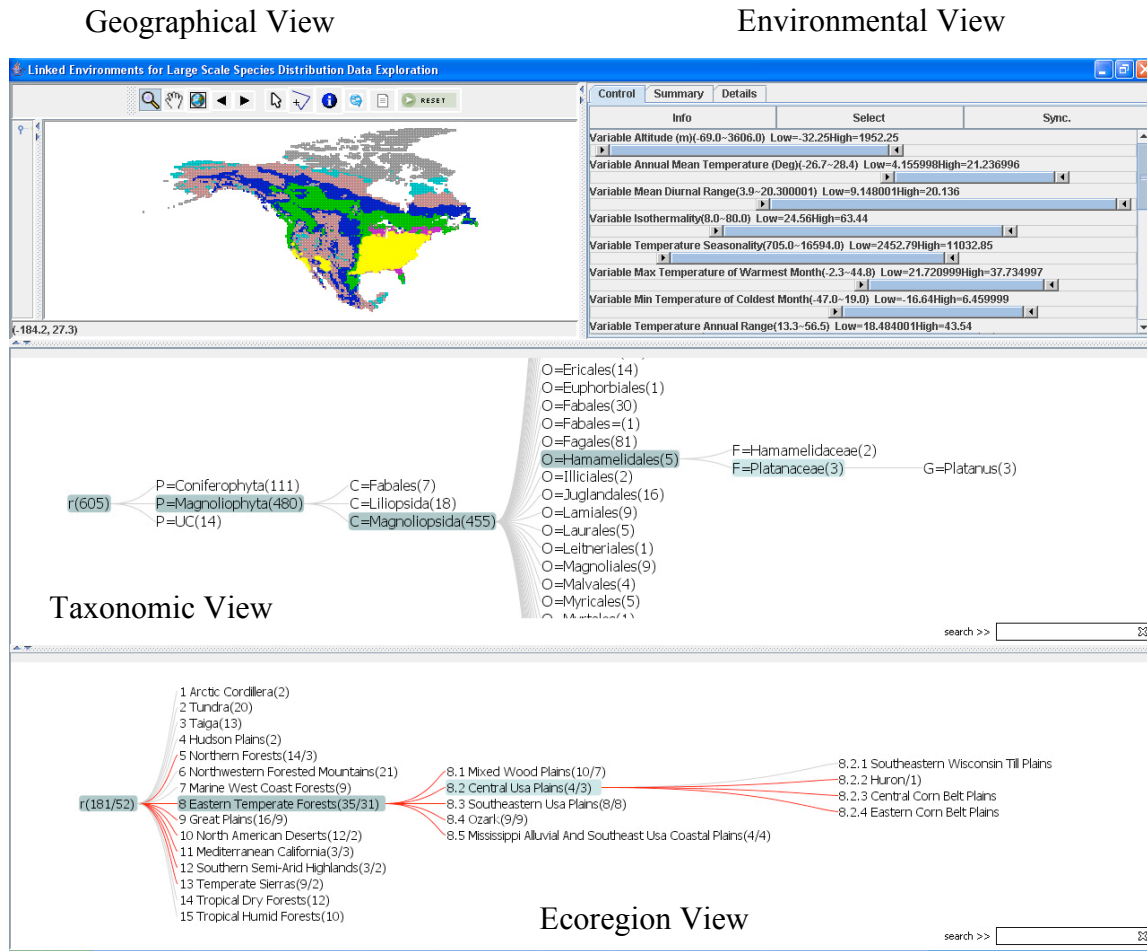


Fig. 1 Screen Snapshot of the four views in the prototype system

The *Geographical* view plays a unique role among the four views in the sense that it is designed to present the distribution information and it displays all cells at the same time to provide users an overview. The subset of cells identified by the operations in other views can be highlighted and contrasted with the rest easily. Furthermore, the *Geographical* view can combine the cells resulting from the operations in the four views. For example, cells corresponding to an ecoregions at a particular level, cells in which a particular taxa is distributed, cells corresponding to a set of environmental envelopes, in addition to cells that intersect with a user-defined environmental gradient. The purpose of the design is to use the *Geographical* view as the “context” and the other views as the “focus” under the “Focus+Contex” visualization framework (Ivan et al. 2000) with respect to species distribution explorations. The *Geographical* view provides standard GIS functions by embedding the JUMP GIS (<http://www.vividsolutions.com/jump/>). LEEASP also extends JUMP GIS to provide additional functions, for example, *Line Intersection* allows users to draw a polyline on a base map. Cells that intersect with the polyline will be selected and any operations for the selected cells can thus be performed. The design is to help ecologists and biogeographers to explore the “gradients” or “transects” of interests. The variations of species compositions along environmental gradients/transects are of particular interest to ecological research (e.g., Willig et al. 2003) but are poorly supported in traditional GIS environments.

The *Environmental* view displays the environmental envelopes of a subset of cells and the values of the environmental variables of the cells. The *Environmental* view has three components (Fig. 2), namely *Control*, *Summary* and *Details*, and they are implemented as tab pages in the prototype system. The *Control* and the *Summary* tab pages are the two presentations of the environmental envelopes of the selected cells. The *Details* panel implements a Parallel Coordinate Plot (PCP, Robert 2003) and a sortable table and they are linked to allow users to look into details of the environmental records corresponding to the selected cells. The *Control* panel, as the name suggests, can also be used to specify the environmental envelopes that users might be interested in and explore the relationship between environmental variables, geographical distributions, the associated taxa and ecoregions.

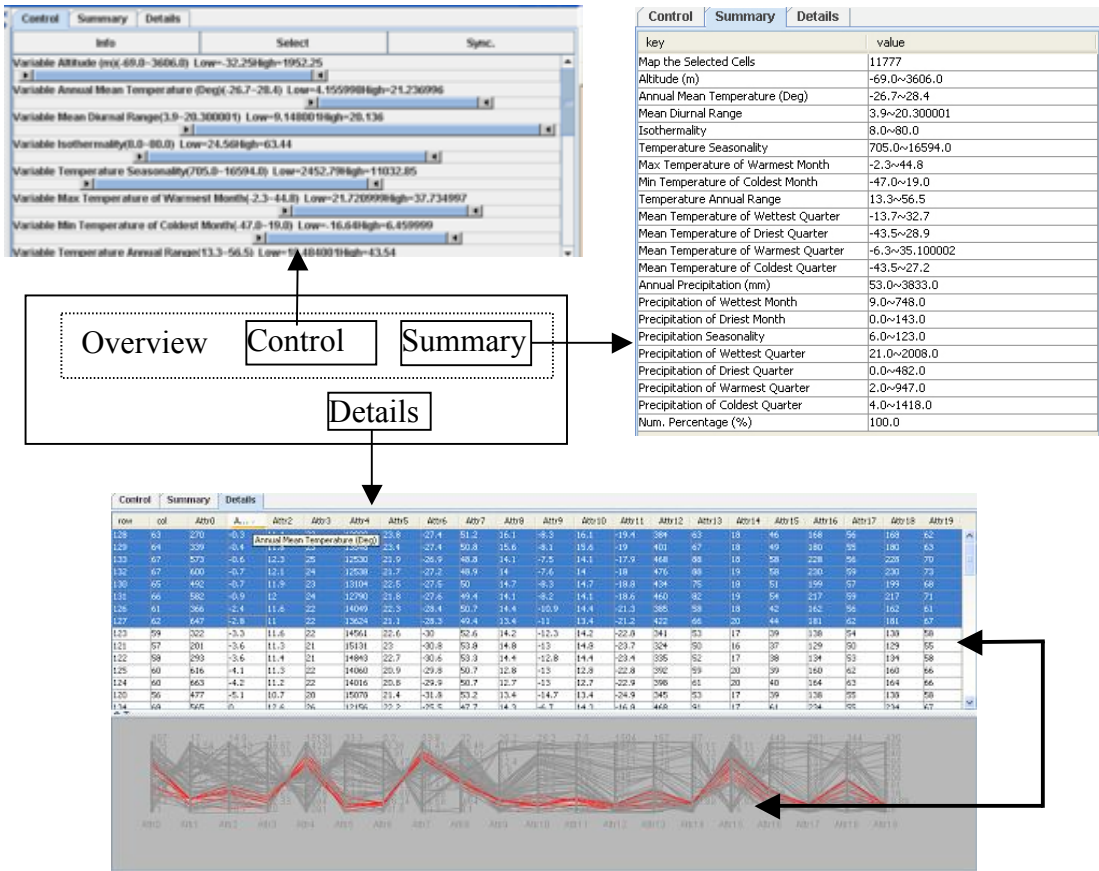


Fig. 2 Components and Coordination in the Environmental Data View

The *Taxonomic* view displays all the species distributed in a region and their higher taxa in a tree format (Fig. 1). In this study, we adopt the Darwin Core (<http://darwincore.calacademy.org/>) and use the following eight levels of taxonomy: Kingdom/ Phylum/ Class/ Order/ Family /Genus/ Species/ SubSpecies. Hereafter we will refer to these eight levels of taxonomy as taxonomic ranks and taxon names at all taxonomic ranks as taxa. As discussed before, we represent the species distributed in a study area as a tree and term it as a taxonomic tree hereafter. We use Prefuse (Jeffrey et al. 2005) to visualize the taxonomic tree and its sub-trees. In the Taxonomic view, tree nodes are labeled like “O=Fabales(30)” where “O=Fabales” is the value of the label field and 30 is the number of species under the taxa. The letter O refers to a higher taxonomic rank, the Order. In other word, the taxonomic rank Order



named Fabales has 30 species. When users work on a non-leaf node (representing higher taxa), all the leaf nodes under the non-leaf node are retrieved on the fly and combined to generate the distribution data of the taxa represented by the non-leaf node. When a subset of cells is selected in the other views (e.g., identifying an environmental gradient/transect by using the *PolyLine Intersection* function), the *Taxonomic* view will identify the paths in the taxonomic tree that correspond to the species distributed in the cells and update labels of tree nodes dynamically. Now the labels of the nodes look like “O=Fabales(30/15)”. The first number in the bracket tells the amount of species (or species richness) under the taxa represented by the node for the whole dataset and the second number tells the amount of species under the taxa for the selected cells (Fig. 1). The ratio of the second number to the first number can be used to tell to what degree the selected cells have the same species richness as the whole study area. We have added a few typical operations supported by Prefuse for trees to the *Taxonomic* view, such as zooming in/out of the canvas, animation when a tree node is expanded and highlighting the nodes along the path from root to the node that the user is currently exploring. When a tree node is expanded (showing the details of the node), other nodes that are the decedents of the sibling nodes of the chosen node will collapse. However, the nodes in the path from the root to the node being chosen and the siblings of the nodes in the path (context) will be kept.

The *Ecoregion* view uses the same tree visualization techniques to visualize the hierarchy of the ecoregions in the study area. The relationship between the cells and ecoregions is one to many, i.e., an ecoregion has many cells of the ecoregion category but a cell can only belong to a single most-detailed (leaf node) ecoregion category. This relationship is simpler than the relationship between the cells and the taxa which is many to many, i.e., a cell can have multiple species and a species can distribute in multiple cells. As a consequence, there will be only a single path from the root to the ecoregion tree node being chosen. For the sake of clarity, LEEASP highlights all paths of the sub-tree rooted at the chosen node (see Fig. 1).

### 3 Related Work and Summary

The work is motivated by the Climate-Vegetation Atlas of the North America project at the USGS in 1990s (<http://pubs.usgs.gov/pp/p1650-a/>). The atlas presents information on the modern relations between climate and the distributions of 407 plant taxa and biogeographic entities from across North America at the 25 km grid resolution. The work is also related to an ongoing project called A Climate Change Atlas for 80 Forest Tree Species of the Eastern United States (<http://www.fs.fed.us/ne/delaware/atlas/>) which delivers tree species distribution maps under different climate change scenarios and the value tables of environmental variables of the species through the Internet.

LEEASP focuses on dynamic visualizations through user interactions. Different from the above two works that deliver static maps and text for individual tree species or predefined species groups, LEEASP allows to select species of all taxon ranks and map their geographical distributions, ecoregion classifications, and envelopes of environmental variables. More generally, the prototype provides multi-way mapping among geographical, ecoregion, environmental, and taxonomic data; and the views representing the four types of data are coordinated. When a subset of data in one view is selected through the graphic user interfaces, by drawing a rectangle on the map, changing the envelopes of environmental variables or by following the paths of taxonomic tree or ecoregion classification tree, the subset of data will be identified in other views so that users can explore the relationships among geography, ecoregion, environmental variables and taxonomic data. In addition, LEEASP provides functionality to

define environmental gradients by on-screen digitization (visually and interactively) and analyze the environmental variable values of the cells by utilizing a variety of visualization techniques. The LEEASP prototype system, including documentation, source codes, binary distributions, third-party libraries and data, is publically available at <http://spica.cs.ucdavis.edu/tech/LEEASPV10.zip>. We encourage interested readers users to try LEEASP by following an easy-to-install process.

### **Acknowledgements**

This work is supported in part by NSF grant ITR #0225665 SEEK and NSF grant ATM #0619139 CEO:P-COMET. We thank Dr. Weimin Xi at TAUM and Anantha M. Prasad at USDA Forest Service for evaluating the prototype and providing constructive suggestions.

### **References**

- Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25(15), pp. 1965-1978. <http://www.worldclim.org/>
- Ivan, H., Guy, M. et al., 2000. Graph Visualization and Navigation in Information Visualization: A Survey. *IEEE Transactions on Visualization and Computer Graphics* 6(1), pp.24-43.
- Jeffrey, H., Stuart, K. C., James, A. L., 2005. Prefuse: a toolkit for interactive information visualization. *Proceedings of the SIGCHI conference on Human factors in computing systems*. Portland, Oregon, USA, ACM Press. Also see <http://www.prefuse.org/>
- Parr, C. S., Lee, B., et al., 2007. EcoLens: Integration and interactive visualization of ecological datasets. *Ecological Informatics* 2(1), pp.61-69.
- Robert M. E., 2003. The parallel coordinate plot in action: design and use for geographic visualization. *Comput. Stat. Data Anal.* 43(4), pp.605-619.
- Willig, M. R., Kaufman, D. M., et al., 2003. Latitudinal gradients of biodiversity: Pattern, process, scale, and synthesis, *Annual Review of Ecology Evolution and Systematics* 34: 273-309.
- Zhang, J., Pennington, D., Liu, X., 2007, GBD-Explorer: Extending Open Source Java GIS for Exploring Ecoregion-Based Biodiversity Data, *Ecological Informatics*, 2(2), pp.94-102.

## Contributed Poster Abstracts

---

### ARCHITECTURAL AND FUNCTIONAL REQUIREMENTS FOR AN ENVIRONMENTAL SENSOR NETWORK COMPUTING PLATFORM FOR TERRESTRIAL BIOLOGICAL RESEARCH STATIONS AND ECOLOGICAL OBSERVATORIES

**Ankit Agarwal, James Beach, Julio Ibarra**

**Abstract** The primary objective of this research is to analyze and derive the requirements for a wireless sensor network software platform for biological research stations. The scope of the analysis includes data management and end-user usability issues such as those associated with data acquisition, validation, reduction, error detection, preprocessing, filtering, formats, caching, alerts, web service and publishing functions. Front-end user interface design requirements are being assessed by using user-centered analysis techniques with existing sensor project leaders and with prospective users, primarily at the Organization for Tropical Studies' La Selva Research Station in north central Costa Rica. We are analyzing researcher needs from a perspective of data processing models and interactions with workflow integration and modeling environments. This work is being accomplished by holding interviews with researchers from both the U.S. and Costa Rica. Also, through workshops and conferences related to sensor networks for environmental monitoring we have gathered science requirements from biologists, researchers, and educators regarding what would aide their research or their educational objectives.

Our preliminary findings have shown that researchers, site systems administrators, and educators require a platform-independent wireless sensor network software platform which would be easy to install and maintain and which would be modular and extensible to local needs. It should be able to accommodate various data types and proprietary protocols emanating from a range of commercially available environmental sensors. The security system should be able to compartmentalize data and set granular permissions for project access and other data streams available. There should be a standards-based schema with robust database query and retrieval mechanism. The database should be able to handle metadata along with environmental data. The architecture should be able to provide data in real-time. There should be web-based visualization and post-processing tools. The web pages should be designed to explain the significance of the data and the relevance of the research to important environmental research issues for technical and non-technical audiences.

### LTER INFORMATION MANAGERS: A COMMUNITY OF PRACTICE

**Karen S. Baker<sup>1</sup>, Nicole E. Kaplan<sup>2</sup>, Inigo San Gil<sup>3</sup>, Margaret O'Brien<sup>4</sup>, Florence Millerand<sup>5</sup>**

<sup>1</sup> University of California, San Diego, CA, 92093-0218, USA

<sup>2</sup> Colorado State University, Fort Collins, CO, 80523-1170, USA

<sup>3</sup> University of New Mexico, Albuquerque, NM, 87111, USA

<sup>4</sup> University of California, Santa Barbara, CA 93116, USA

<sup>5</sup> Universite du Quebec a Montreal (UQAM), Montreal, QC H3C 3P8, CANADA

**Abstract:** Communities of Practice are groups of people who share a concern or a passion for something they do and who want to learn more about how they do it. Such a community is more than a group of people having the same job or a network of connections between people. Three elements characterize a Community of Practice: 1) the domain, 2) the community, and 3) the practice. Regular interaction such as with an annual meeting is a key integrative mechanism that brings into play elements of practice including agenda setting, knowledge management, professional development, advocacy, and resource mobilization. The history and multi-dimensional aspects of Communities of Practice provide a framework for considering information management organizationally through structures that facilitate communication and learning. We explore the Long Term Ecological Research Information Management Committee in particular as a Community of Practice. Examples of how the information management role has emerged and is defined within the Long Term Ecological Research community will be presented. How the committee as a collective fits within this framework will be considered by taking into account interests, activities, and relations. Active membership, professional engagement, and collective learning are needed to ensure relevance as well as long-term sustainability.

### **RECENT INFORMATION MANAGEMENT SYSTEM ENHANCEMENTS AT THE NORTH TEMPERATE LAKES LTER**

**David Balsiger, Barbara Benson, Jeff Maxted, Luke Winslow**

**Abstract** The information management team at the North Temperate Lakes (NTL) LTER has been developing enhanced functionality in several areas: data acquisition, data access including expanded access to spatial data, and the management of sensor network data. Two in-house programs for data acquisition have been substantially modified. MobileFish is used by the crew collecting annual fish data and provides data capture on a PDA. Key features of this application include the capability to set up sampling events prior to going into the field, the efficient entry of data and metadata during the actual sampling, prompting by the software that promotes adherence to sampling protocols, and a built-in data screening algorithm for fish lengths and weights. Z3, a program originally designed for zooplankton counting and measuring, is being reworked to include extra features and to be customizable for other counting and measuring uses such as those with benthic invertebrates or fish scales. The management of sensor data has challenged us to investigate new data models and collaborate to develop new tools. The Vega data model is a flexible database architecture designed to accommodate additions and changes in sensor deployments without database structure changes. GLEONDN has been developed to handle simple quality assurance/quality control, exceptions in streaming data, and insertion of data into a repository. dbBadger is a web-based application and allow users to quickly and easily discover and retrieve stored sensor data. dbBadger can also align, interpolate, and aggregate time series data based on the needs of the user. To enhance data discovery we have designed and are beta-testing a search interface to the NTL Data Catalog on the NTL website. This interface allows the user to select a Project type, Theme, Location, and Period of Interest as well as to add search criteria for text strings in various metadata fields such as dataset title, investigator or species name. Migrating to a server-based GIS architecture is enabling us to provide spatial data through web mapping services. This enhances our capacity to serve updated vector and raster data to users from our extensive spatial data catalog. Server-based GIS enables us to serve geoprocessing services, including ecological models that users can apply to spatial data via an internet browser.

## IMPROVING METADATA SEARCH EFFICIENCY BY ENABLING SEMANTIC QUERIES

**Chad W Berkley, Shawn Bowers, Matthew B. Jones, Mark Schildhauer**

**Abstract** Increasing amounts of digital ecological data are becoming available (e.g., over 15,000 datasets in the Knowledge Network for Biocomplexity alone), making it critically important to improve techniques for more precisely locating and delivering relevant information from scientific searches of these resources. Semantic technologies hold the promise of enabling powerful "smart" search of online data archives. Here we describe how we are constructing semantic search features within the Metacat XML database system, which is used by many ecological research sites around the world for archiving their data using a standardized metadata format. The prototype semantic search system in Metacat uses a system of OWL-DL ontologies, such that ontological concepts can be linked to specific features and attributes of the Metacat data holdings, via an XML-based annotation language. Queries are then resolved through a free, widely-available reasoning engine that can yield effective search results due to leveraging the ontological structures. We have architected Metacat to seamlessly store and access ontologies alongside the datasets and their associated annotations and metadata, making it easy for any Metacat implementation to harness the power of semantic queries. In the future as data repositories continue to grow, these tools will be instrumental in helping scientists locate and interpret data for their research needs.

## ENVIRONMENTAL DATA UPLOAD AND VISUALIZATION TOOLS

**Shira Bezalel**

San Francisco Estuary Institute

**Abstract** Easy access to reliable data is a primary objective of any environmental information management system. Providing high quality, scientific information allows for the formulation of technically sound policies and the ability to address specific management questions. Tools can assist with the flow of information thorough the various data management steps of data collection, uploading, and review, and facilitate the retrieval, exchange, and visualization of results. This poster highlights tool development from two projects managed by the San Francisco Estuary Institute. The Regional Monitoring Program for Water Quality is the primary source of long-term contaminant monitoring data for the San Francisco Estuary and annually collects water, sediment, and tissue samples. The South Bay Mercury Project is a collaborative, three-year project that characterizes mercury in the sediment, water, and biota ("sentinel" species) indicative of different landscape management endpoints in the San Francisco Estuary's South Baylands.

Field data collection entry forms have been developed in Access for both of these projects and enables data to be easily uploaded to the main database. These entry forms have saved in both staff time and costs required for entering standardized information into project databases. Constraints within the form prevent entry of erroneous data by forcing users to select from pre-defined code lists, such as analyte and device names.

The South Bay Mercury Project uses Google Earth as part of its visualization tool for reporting mercury results at specific sample sites. Concentrations are differentiated through a range of colors and symbol heights. This tool provides scientists with an essential aerial perspective in which to evaluate the results.

The Regional Monitoring Program makes its 14-year dataset available online through a user-defined query tool, from which results can be downloaded into an Excel file in two formats: cross-tabulated, making it easier for reviewing data across stations and time, and flat-file, for importing data into statistical programs. The development team is currently involved in two major enhancements to the query tool: the ability to download data from other projects collecting data in the Estuary and the availability of a visualization tool for dynamic mapping of concentrations. These enhancements will facilitate the mapping of results to show the spatial distribution across the entire Estuary.

Tools assist with the uploading and standardizing of data and increase the options for reporting data to the scientific community and general public in a more meaningful way that address specific management and research questions.

## **AN INTEGRATED FRAMEWORK FOR HYBRID AND ADAPTIVE MODELING OF SEA SURFACE TEMPERATURE: A WORKFLOW-BASED APPROACH TO COMPARISON**

**Daniel Crawl, Peter Cornillon, Ilkay Altintas, Nathan Potter, James Gallagher, Mark Schildhauer, Matthew B. Jones**

**Abstract** Sea surface temperature (SST) fields are among the most broadly used observational data sets related to the ocean, and constitute critical information for informing a broad range of analyses and models, ranging from estimates of near-surface currents and water body masses, to application in biodiversity models, support of search and rescue missions, as well as the investigation of air-sea interaction at many scales. There is a bewildering array of SST products available, many deriving from satellite-borne instruments, as well as ship-board and other in situ instruments. Quantitative comparison and integration of these various SST data sources is currently extremely difficult and time-consuming.

This poster presents a case study to develop Kepler scientific workflows to facilitate the quantitative evaluation of SST data sets. The presented workflow is comprised of three main steps, namely, a user input sub-workflow, a match-up generation sub-workflow, and a statistical analysis sub-workflow. The user input sub-workflow provides the user with an interface to specify how the match-up database is to be constructed and which SST data sets are to be compared. The match-up generation sub-workflow produces a match-up database from the selected SST datasets. Finally, the analysis sub-workflow performs a suite of statistical analyses on the match-up database. The workflow generates a KML file based on the results of this analysis that can be displayed using Google Earth.

The presented work is part of a National Science Foundation (NSF) funded project called Realtime Environment for Analytical Processing. (REAP, <http://reap.ecoinformatics.org>).

## **ELECTRONIC COLLECTION OF VEGETATION MAPPING DATA WITHIN THE GRAND CANYON NATIONAL PARK**

**Scott C. Curran, Mike Kearsley**

Grand Canyon National Park

**Abstract** Working 8-24 days at a time in extreme remote locations within the boundary of the national park, research crews needed an electronic data collection process that would endure harsh elements as well as ensure that data was collected properly. The equipment that was used had to be capable of sustaining long battery charges, lightweight, durable and remain relatively inexpensive. The software application used to record the data needed embedded quality controls and also safe guards to protect the data from being lost. Each crew was given 2-3 handheld data recorders and one semi-ruggedized ultramobile pc. Data was recorded in Pendragon Forms 5.1 software on the handheld data recorders. At the end of each day, all handheld data recorders were synchronized to an Access database on a semi-ruggedized ultramobile pc. After synchronizing all the handhelds, the Access database was then copied to a ruggedized USB flash drive for a daily backup. The data interface on the Pendragon Forms 5.1 application was intentionally kept simple in design; so, the recording process would be more transparent to the user. During the application design phase crews tested the application on several occasions to provide feedback, but this testing also gave the crews a chance for hands-on learning. This data was collected electronically over the course of several trips this spring and summer. The crews eventually became comfortable using the electronic data collection process and no data was reported lost. There were some problems with poor contrast on the handheld data recorder screen when used in direct sunlight, but nothing that could be resolved. Overall, the process was successful for this initial recording period, but future documentation should help new users adapt more quickly to this electronic data collection process.

## **CROSS-SITE ANALYSIS OF ABIOTIC DRIVERS AND ANPP AT FIVE GRASSLANDS SITES**

**Judith B. Cushing, Nicole E. Kaplan, Christine Laney, Carri LeRoy, Juli Mallett, Ken Ramsey, Kristin Vanderbilt, Lee Zeman**

**Abstract** The Grasslands Data Integration Project collects data on Aboveground Net Primary Productivity (ANPP) from five grasslands study sites that encompass a variety of different ecosystems into a centralized database containing species and growth form information at the individual plot level. Data was integrated across varying species protocols, experimental designs, and collection methods without aggregation or data loss. The resulting observation-centric database supports statistical aggregation by a variety of factors including species, family, growth form, vegetation biome type, and physical location. We have done preliminary ANPP comparison between sites as well as cross-site analysis of the roles of abiotic drivers such as temperature and precipitation on ANPP, and the role of prominent species such as yucca elata. This poster presents our database schema and some preliminary analysis of the influence of abiotic drivers, including the Palmer Drought Severity Index, maximum and mean daily temperatures, and mean precipitation.

## **PROBLEMS AND SOLUTIONS IN SPECIES-CODED DATA: BEST PRACTICES AND COMMON ISSUES**

**Judith B Cushing, Juli Mallett, Lee Zeman, Nicole Kaplan, Christine Laney, Ken Ramsey, Kristin Vanderbilt**

**Abstract** Ecologists are interested in conducting cross-site or large-scale integration and analysis of annual aboveground net primary productivity (ANPP) values, but are often hindered by the lack of standard methodologies for data collection, data management practices and detailed metadata documentation across sites. The Grasslands ANPP Data Integration (GDI) project has brought together experts in ecology, information management, and computer science to address the challenges of integrating ANPP data and create a centralized database. The integration of species-coded data between sites proved to be a major component of the work and revealed a number of common problems. Based on our experience, we have developed a number of suggestions for managers of any species-coded database to minimize the problems of cross-site integration and concise examples that convey the problems facing all integrators of species-coded data. This poster suggests guidelines for species data formats, reclassifying species or combining indistinct species, correlating species lists across sites, and referencing standard species lists such as the US Department of Agriculture's PLANTS database. These techniques are especially important for long-term, ongoing, or integrated datasets.

## **SIMPLIFIED DEPLOYMENT OF THE METACAT DATA AND METADATA SYSTEM**

**Michael Daigle, Matthew B. Jones, Benjamin Leinfelder, Shaun Walbridge, Jing Tao**  
National Center for Ecological Analysis and Synthesis, University of California Santa Barbara  
**Abstract** Most data management systems used in environmental field stations are custom systems that integrate many software components to serve the needs of that station. Although focused on the needs of these organizations, custom-developed systems require significant expertise in information technology and software development in order to be successful. Many smaller field stations, laboratories, and individual researchers lack the resources and technical expertise to develop their own custom system. We have extended the Metacat data and metadata catalog system to provide a more simplified system for installing and deploying a data management system with minimal technical expertise. The first phase of this project has focused on simplifying the configuration of the system to make it approachable by researchers. The second phase will focus on installing system prerequisites in order to make the overall system installation process quick and straightforward on multiple computing platforms. By creating this turnkey data management system, we hope to increase the number of laboratories and researchers that utilize a structured metadata and data management system and thereby have a significant impact on the accessibility of ecological and environmental data.



## THE NORTH AMERICAN CARBON PROGRAM GOOGLE EARTH COLLECTION

**Peter C. Griffith, Lisa E. Wilcox, Amy L. Morrell**

Science Systems & Applications, Inc. and the Carbon Cycle and Ecosystems Office, NASA  
Goddard Space Flight Center

**Abstract** The Google Earth Collection for the North American Carbon Program provides a central geospatial data search and discovery tool that allows scientists and agency program managers to browse research products being contributed by each NACP project to the synthesis efforts addressing the core questions of the Program.

## MANAGING INFORMATION FOR ENVIRONMENTAL FLOWS IN TEXAS

**Eric S Hersh, David R Maidment**

Center for Research in Water Resources, The University of Texas at Austin, USA

**Abstract** Environmental flow is water left in or released into a river system, often for managing some aspect of its conditions. The goal may be, for example, the broad maintenance of a healthy river ecosystem or the narrow focus of ensuring the survival of an individual species. Relevant data describing the stream flow, water chemistry, geomorphology, and biology of streams and rivers is often contained in a variety of formats and in many geographic locations. Thus, an information system is developed to organize and make available data relevant to the study of environmental flows in a consistent and accessible format. Relevant data from hydrology and hydraulics, water quality, climatology, geomorphology and physical processes, and biology is assembled to facilitate data discovery, acquisition, and sharing.

Working symbiotically with the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic Information System (HIS) project, an NSF-supported effort to improve access to hydrologic data at the nation's universities, environmental flows data is stored in the CUAHSI Observations Data Model and web services are established for the computer-to-computer communication of data in order to extract data from disparate sources in disparate formats, to transform the data into the common language of CUAHSI WaterML, and to load the data into an end user's system. The environmental flows information system includes a linkage to a georeferenced digital archive of documents providing for parallel access to both data and the knowledge products derived from that data. Via the Data Model and accompanying Document Model, an information system capable of managing observational data, geographic data, modeled/constructed data, and documents is offered.

A prototype environmental flows information system is developed for the State of Texas which incorporates relevant known available datasets from federal, state, academic, river basin, and local sources. Tools are developed to assist in the publishing, visualization, and access of data and documents via map-based, spreadsheet-based, and other methods. The information system might be used to provide: (1) rapid low-cost data integration, (2) improved data access for the public, and (3) support for the analysis and determination of environmental flow needs. The environmental flows information system represents the integration of the physical, chemical, and biological information for rivers and streams in a consistent and accessible manner in one system in one place.

## **USGS NBII RELEASES RE-DESIGNED INTERFACE FOR METADATA CLEARINGHOUSE**

**Vivian Hutchison**

**Abstract** The use of standards for data documentation allow scientists and researchers to discover completed or on-going research projects in a particular area of study, which can lead to new opportunities for collaboration and data sharing. The National Biological Information Infrastructure (NBII) maintains various standards to support data interoperability. In preparation for the release of ISO 19115 North American Profile, the NBII has released a new, enhanced version of the NBII Clearinghouse. With powerful search capabilities and updated features, users can search geographically or by specifying particular data providers, then bookmark or email record results. An RSS feed can be set up to inform a user about new records in the Clearinghouse reflecting a particular query. The Clearinghouse is supported by Mercury technology through the Oak Ridge National Laboratory. Visit the Clearinghouse: <http://mercury.ornl.gov/nbii>. A poster will be presented highlighting the features of the new NBII Clearinghouse capabilities.

## **A TEAM APPROACH TO DATA SYNTHESIS: THE PLAYBOOK FOR CREATING A CENTRALIZED, DYNAMIC, AND SUSTAINABLE ANPP DATABASE**

**Nicole E. Kaplan, Kristin L. Vanderbilt, Lee Zeman, Judy B. Cushing, Christine Laney,  
Juli Mallett, Ken Ramsey, Jincheng Gao, Judith Kruger, Carri LeRoy, Daniel Milchunas,  
Esteban Muldavin**

**Abstract** The Grasslands Data Integration (GDI) project has brought together ecologists, information managers and computer scientists to address the interdisciplinary challenges of integrating aboveground net primary productivity (ANPP) data from multiple LTER sites to facilitate analysis of this important ecosystem variable. Each LTER site uses a different experimental design for collecting the data and employs site-specific naming conventions for the plant species studied. Development of the relational database to house the integrated data product required that ecologists and computer and information specialists work closely together to resolve semantic issues and to ensure that the data accessible from the database would be at a standard resolution for all sites. The integrated database has been used to generate some preliminary analyses relating ANPP to meteorological data that clearly demonstrate the value of this tool for the scientific community. Future plans for the GDI include establishing an ANPP data warehouse, developing a user friendly browser, and automatically generating QAQC reports for newly ingested data. The GDI collaboration is an example of how professionals with inter-related work experience build a community of experts and a successful data product for the LTER (Baker and Millerand 2007).

## **ABSTRACTING FUNCTIONALITY AND ACCESS: FACILITATING DATA SYSTEM MANAGEABILITY AND SITE COORDINATION**

**Mason A. Kortz, James E. Connors, Karen S Baker**

Scripps Institution of Oceanography, University of California, San Diego, 92093-0218, USA

**Abstract** As the functionality of site data systems increases, frequently so does the complexity. Organizing system functionality through distinct layers of abstraction, from low-level system access to high-level user access, is key to maintaining a manageable system. Toward this end, a data system that is an interdependent set of databases, files, and other resources can often be abstracted into a relatively compact set of data access methods. Abstraction layers allow developers to leverage not only the content of a data system but the organizational logic as well. Leveraging may take the form of facilitating local site reuse or sharing across projects and sites. Abstraction enables the development of multiple applications, accessing the same data system - and its data - via a single interface layer. This poster explores three models by which data access methods may be abstracted and shared: application programming interfaces, remote procedure calls, and resource state transfers. Each model is defined in general as well as illustrated by examples designed, developed, and deployed at two Long-Term Ecological Research sites (Palmer Station and California Current Ecosystem).

## **NORTH OF IRELAND COASTAL MONITORING PROGRAMME - QA FOR AN OPERATIONAL NETWORK OF MOORED OCEANOGRAPHIC INSTRUMENTS**

**Adam Mellor**

**Abstract** An operational network of 12 remote monitoring stations around the Northern Irish coastline play a sentinel role in water quality monitoring. The programme provides near-real time, high resolution data for water quality in the coastal zone and allows the capacity for reactive management to complement routine monitoring surveys. This network has to the current day disseminated (and quality assured) data on an ad-hoc basis - data from remote in-situ moored instrumentation is rarely without flaw as influences such as fouling and drift compromise measurements. Objective near-real-time quality assurance (QA) is being developed retrospectively with consideration to the demands of policy based criteria such as the concentrations of nitrates, dissolved oxygen and chlorophyll from the European water framework directive (WFD). Data will be automatically integrated with targeted QA reference measurements allowing them to be validated & filtered for rapid dissemination with quantifiable and appropriate levels of confidence.

An analysis of survey and time series data prioritises sites for continuous monitoring, and characterises the relationships between sites to give enhanced confidence in the value of fixed point monitoring with moored instrumentation. This analysis also enables the appropriate design of complimentary water quality surveys using techniques such as optimum allocation analysis (OAA).

Outputs of the moored network in addition to the obligatory monitoring requirements include projects such as the integration of the environmental data into ecosystem modelling programmes such as the SMILE model (Sustainable Mariculture in northern Irish Lough Ecosystems). Now operational, the SMILE model can estimate carrying capacities for aquaculture as well as

forecasting the potential impacts from overexploitation by integrating shellfish growth, hydrodynamic and ecosystem models.

## **BUILDING THE FRAMEWORK FOR A VIRTUAL DATA CENTER FOR ECOLOGY AND THE ENVIRONMENTAL SCIENCES**

**William Michener, Suzie Allard, Paul Allen, Peter Buneman, Randy Butler, John Cobb, Robert Cook, Patricia Cruse, Bruce Dancik, Ewa Deelman, David DeRoure, Mindy Destro, Cliff Duke, Charles Fox, Mike Frame, Stephanie Hampton, Carole Goble, Nancy Grimm, Donald Hobern, Peter Honeyman, Jeffery Horsburgh, Vivian Hutchison, Matthew B. Jones, Steve Kelling, Jeremy Kranowitz, John Kunze, Hilmar Lapp, David Leslie, Jr., Bertram Ludaescher, Thomas Moritz, Lorraine Normore, Robert Peet, Ricardo Pereira, Line Pouchard, Jim Reichman, Hannu Saarenmaa, Robert Sandusky, Ryan Scherle, Mark Schildhauer, Mark Servilla, Kathleen Smith, Carol Tenopir, Paul Uhler, Dave Vieglais, Todd Vision, Jake Weltzin, Von Welch, Bruce Wilson**

**Abstract** Data centers (also referred to as data archives or data repositories) have been created to preserve data and explanatory documentation (i.e., metadata), support discovery of data by searching (e.g., by location, time, taxa, keywords), enable data access, and sometimes data processing and integration with other data. In addition to their core functions, some data centers provide other services such as research and development, help-desk support, training, and outreach. Data centers are vital to science because they can provide secure and permanent repositories for the data and information that are legacies of the scientific enterprise, and they can facilitate new research and synthesis efforts.

Despite the proliferation of data centers throughout science, the discovery, acquisition, and integration of the disparate data needed to address the grand environmental challenges are exceptionally difficult, time-consuming, and expensive to achieve. Reasons for this include insufficient metadata, heterogeneity of data and metadata standards, lack of interoperability solutions across data centers, organizational and funding instabilities, and a poorly developed scientific culture of data sharing and data stewardship.

Because many key needs are not presently being met, we propose a new type of organization--a virtual data center--that can bind together existing data centers and provide seamless and straightforward discovery and access to the broad array of data, information, and analytical resources needed to address current and emerging scientific challenges. Steps involved in the formation of such a center, including principles that should guide its organization, required functionality, opportunities for leveraging existing cyberinfrastructure, and potential funding mechanisms are presented. This poster highlights results from a series of data workshops hosted by the Ecological Society of America and supported by the NSF, as well as three proposed implementation efforts (Dryad, INTEROP, and DataNet).

## **IMPLEMENTING AN AUTOMATED PROCESSING SYSTEM FOR LOW-FREQUENCY STREAMING DATA USING AN ECLECTIC APPROACH**

**John H. Porter**

**Abstract** The path streaming data follows from the sensor to a dataset or graph on the World-Wide Web has many steps including ingestion, quality assurance, archival storage, and generation of products for display and download. The software available for accomplishing these steps are widely varied, each with its own strengths and weaknesses. However, no piece of software is best at everything (although many have overlapping capabilities). For this reason, the Virginia Coast Reserve Long-Term Ecological Research Project has developed fully-automated systems for processing low-frequency ( $> 0.10$  hours per measurement) data that build on the strengths of an eclectic mix of software products and computer systems. This poster will provide an overview of a system used to collect and process data from a small (10 node) network of water level recorders located on a Virginia barrier island. Serial and Internet Protocol wireless networks are used to harvest hourly data from Campbell Scientific data loggers, using proprietary Loggernet software that runs on a PC at the Anheuser-Busch Coastal Research Center. Every few hours, Windows scheduler is used to run a batch file that copies the downloaded files to a network-accessible directory on a Unix computer at the University of Virginia. There the Statistical Analysis System (SAS) is used to integrate new data with existing data, including elimination of duplicates, data format conversions (e.g., dates and times) into standard forms, flagging of out-of-range values, and production of an integrated dataset for download by users. That integrated dataset is also used as input to "R" programs on a Linux-based web server to produce a variety of graphical and textual statistical summaries that are automatically posted on the WWW. The advantages of these types of systems are that they require relatively simple programming, each software product is doing what it does best with no need for esoteric programs; that they can incorporate a variety of computers and operating systems, taking full advantage of what is available; and finally, that they can operate unattended for months at a time, reliably providing data to users with minimal operator intervention.

## **A FRAMEWORK FOR DEFINING AND ENFORCING MULTIPLE VALIDATION ENVIRONMENTS (I.E. PROTOCOLS) WITHIN AQUATIC ECOLOGY**

**Steve Rentmeester**

**Abstract** The listing of Pacific salmon under the Endangered Species Act led to an evaluation of how natural resource agencies in the Pacific Northwest collect, valid, analysis, report and share aquatic resources data. In May 2000, the Independent Science Review Panel (ISRP) released a report documenting the inadequacies of the data management system in the Columbia River Basin and noted significant inconsistencies in how aquatic resources data were reported. In 2007, Environmental Data Services developed the Aquatic Resource Framework (ARF) to support agencies enter, valid, document, and analyze aquatic resources data. The framework assumes there is a finite list of real world objects and attributes that are relevant to decision making about aquatic resources and assumes an infinite number of protocols that describe how these real world objects and attributes are observed or measured by data collectors. Each protocol is stored as unique data dictionary within the framework. These data dictionaries are then called by the front-end application to define the validation environment for a given protocol. Data entry sessions are associated with a single protocol and protocol-specific validation is enforced during data entry.

Use of a finite ontology of real world objects supports documentation, integration, discovery, analysis, and sharing of data across multiple survey types (e.g. smolt trapping survey, water quality survey, stream habitat survey, snorkel survey, electro-fishing survey, etc.) Examples of real world objects include Fish, Stream, Large Woody Debris, Habitat Unit, Transect, Station, and Channel Segment.

### **PROMOTING COMMUNITY CONTRIBUTIONS WITH HIGHLY CONFIGURABLE COMPONENT BASED SOFTWARE, A KEPLER ARCHITECTURE**

**Aaron Schultz, Matthew B. Jones, Timothy McPhillips, Sean Riddle, David Welker**

**Abstract** Kepler is a system that allows scientists to utilize a wide variety of data stores and analysis tools from many disciplines in an integrated software system through the metaphor of a scientific workflow. The large number and variety of tools used in environmental information management systems makes the interconnection and ease of use of these tools a difficult and daunting task. A standardized way for modularizing, packaging, interconnecting, and managing these tools is needed. We have designed a new architecture for Kepler that incorporates the Open Services Gateway Interface standard, a well established and proven software framework, into the Kepler Scientific Workflow system, allowing for highly configurable, component based interoperability of a large and growing number of diverse software systems in a standardized way.

### **EARTHGRID WEB SERVICES FOR ACCESSING HETEROGENEOUS DATA SYSTEMS**

**Jing Tao, Matthew B. Jones, David Vieglais, Arcot Rajasekar, Lucas Gilbert, Benjamin Leinfelder**

**Abstract** EarthGrid services are web services which provide a high-level programmatic interface for existing data and computer services. It is a lightweight layer that can be implemented easily to expose data from existing systems using a shared web service interface. Any data system wrapped by the EarthGrid API can be accessed by a common client, making it easier for client systems to access multiple data servers. Currently, several data management systems have implemented the EarthGrid interface, including Metacat for ecological and environmental data and metadata, Digir for natural history collection specimen data, and SRB for scientific data from several disciplines. Each of these systems has implemented EarthGrid interfaces for metadata query and data retrieval, and some have also implemented the interfaces for user authentication, data deposition, and other advanced services. The EarthGrid provides a registry for systems that have implemented one or more services, making it easy for clients to discover data systems that are available through the EarthGrid. The Kepler scientific workflow system is an example client that uses the EarthGrid to seamlessly and transparently access multiple heterogeneous data systems for analyses and simulation.

## **INTEGRATING GOOGLE EARTH AND INTERNET MAPPING INTO YOUR WEBSITE**

**Theresa Valentine**

**Abstract** Researchers on the HJ Andrews Experimental Forest Long Term Ecological Research (LTER) Site have been interested in visualizing their data using Google Earth technology. ESRI's ArcServer architecture provides a framework for developing high quality internet mapping applications that have the option of being enabled to work with Google Earth. Internet applications can then function as stand-alone applications and be integrated with Google Earth on the user's desktop. In addition, Google has recently released an application that imbeds Google Earth into a webpage, so the user does not need to download and start up Google Earth. They will need to download a small application (one time download) that will start when the web page is accessed.

This technology provides an easy to use interface with 2005 high quality color digital orthophotography (available for Oregon). The user can zoom in and out, pan, change perspective to fly through their area of interest, and add their data to the application. ArcServer services can be modified to provide subsets of data to Google Earth, and users can add the network links to their desktop. The core data is stored in the spatial data depository and the user moves seamlessly between GIS and Google Earth.

## **WEB-BASED COLLABORATION IN AN ECOLOGY THINK-TANK**

**Shaun Walbridge, Mark Schildhauer, Jim Regetz, Matthew B. Jones, Rick Reeves**

**Abstract** Rising costs of transportation, inherent limitations of email communication and the ad hoc collaboration methods typically employed in virtual meetings has led NCEAS to implement software solutions that provide integrated mechanisms for working together remotely. Focusing on the current needs of working group participants we have found a mix of easy to use technology that facilitates and encourages collaboration in key areas such as: group discussion, literature review, document sharing and event scheduling. By providing low-barrier mechanisms to store documents, data and communications, groups are able to use the tools throughout the lifecycle of their research. Participants have access to important items including initial literature, ephemeral documents, interim results and drafts, and finished products. Capturing metadata and data within the Metacat system is a planned extension that will provide searchable, durable storage for valuable data products.

## **INFORMATION INFRASTRUCTURE: EMERGENT ROLES, RESPONSIBILITIES AND PRACTICES**

**Lynn R. Yarmey, Karen S. Baker**

**Abstract** Human activities together with technical elements and collective practices are core elements for growing local infrastructure as well as for bridging with other communities and networks. Site information management activities create a shared data curation experience where data curation refers to managing the capture, use and preservation of the data. Identifying and elaborating upon local data activities opens up the complex set of arrangements that comprise site information management, including the variety of roles emerging to address mediation and collaboration. Any one activity may be carried out in practice by different participants at each site. That is, what one site considers an information management role may be carried out by a researcher, technician, analyst, or education coordinator at another site. The diverse distributions of responsibilities at each site are a result of meeting local scientific needs with a mix of local participants and practices. Comparing and contrasting different site infrastructure arrangements prompts discussion that deepens our understanding of data and data curation. Insight into data activities and their associated roles and responsibilities may be seen as a preparatory step for conscientiously designing an effective data network.