Proceedings of the

Environmental Information Management Conference 2011 (EIM 2011)




M.B. Jones,
C. Gries, Editors

Proceedings of the Environmental Information Management Conference 2011 (EIM 2011)

September 28-29, 2011
Santa Barbara, CA

Editors:
Matthew B. Jones
National Center for Ecological Analysis and Synthesis (NCEAS)
UC Santa Barbara

Corinna Gries
Long-term Ecological Research Network
University of Wisconsin

Publisher: University of California
Published: 2011
doi:10.5060/D2NC5Z4X

TABLE OF CONTENTS

PROGRAM COMMITTEE

Conference Co-Chairs

Matthew B. Jones
Corinna Gries

Program Committee

Paul Allen
Ilkay Altintas
Ioannis Athanasiadis
Terry Benzel
Luis Bermudez
Kenneth Chiu
Helen Conover
Patricia Cruse
Judy Cushing
Peter Fox
Peter Griffith
Don Henshaw
Jeff Horsburgh
Vivian Hutchison
William Michener
Margaret O'Brien
Eamonn O'Tuama
Richard Pyle
Ryan Scherle
Mark Schildhauer
Wade Sheldon
Karen Stocks
David Tarboton
Sameer Tilak
Dave Vieglais
Jonathan Walsh
Bruce Wilson

SPONSORS

National Center for Ecological Analysis and Synthesis
University of California, Santa Barbara

Long-term Ecological Research Network
University of Wisconsin

DataONE

Environmental Information Management 2011

Wednesday, September 28, 2011

7:00 AM	**Registration and Coffee**	
8:15 AM	**Welcome to EIM** (Santa Ynez Room)	
8:30-10:30 AM	**S1: Sensors and Workflows** (Santa Ynez Room)	**BoF1: Birds-of-A-Feather** (Fiesta Room) Internet mapping: What are the options?
10:30 - 11:00 AM	**BREAK**	
11:00AM-12:00 PM	**Keynote: Patricia Cruse** (Santa Ynez Room)	
12:00 - 1:30 PM	**LUNCH** (on your own)	
1:30 - 3:50 PM	**S2: Semantics and Data Management** (Santa Ynez Room)	**BoF2: Birds-of-A-Feather** (Fiesta Room) 1:30 – 2:30 PM: Using Web Tools and Methods… 2:30 – 3:30 PM: Geospatial Data Management …
4:00 - 7:00 PM	**Poster Session and Reception** (Sierra Madre North Room)	

Thursday, September 29, 2011

8:00 - 8:30 AM	**Coffee**	
8:30-10:30 AM	**S3: Discovery, Visualization, and Analysis** (Reagan Room)	**BoF3: Birds-of-A-Feather** (Fiesta Room) Automating Data Processing and Quality Control…
10:30 - 11:00 AM	**BREAK**	
11:00AM-12:00 PM	**Keynote: Ned Gardiner** (Reagan Room)	
12:00 - 1:30 PM	**LUNCH** (on your own)	
1:30 - 3:50 PM	**S4: CyberInfrastructure Systems** (Reagan Room)	**BoF4: Birds-of-A-Feather** (Fiesta Room) Functional Requirements for the EML Dataset Congruency Checker
3:30 - 4:00 PM	**BREAK**	
4:00 PM - 5:00 PM	**Plenary Discussion:** Community Standards and Practices (Reagan Room)	
5:00 PM	**Concluding remarks**	

KEYNOTE SPEAKERS

I. PATRICIA CRUSE

Building Communities, Partnerships, Tools, and Services in Order to Thrive in a Dynamic Information Landscape

Digital information is vital to the research, teaching, and learning mission of academia. However, technical transformations in research, teaching, and learning; adaption of a more business like model for running the institution; decreased budgets; and emergent trends in the information, search, and publishing industries are all creating major changes in today's research institutions. In addition, the digital environment has fundamentally transformed the way in which information is produced and disseminated within the university, blurring the lines between knowledge creation and formal publication; changing the way users find, access, and use information; and creating new demands for the effective curation of digital content.

In order to respond effectively to these challenges the UC system established the UC Curation Center (UC3) at the California Digital Library (CDL). UC3 is a creative partnership bringing together the expertise and resources of the CDL, the ten UC campuses, and the broader international curation community. We foster collaborative analysis and solutions to ensure the long-term viability and usability of curated digital content. The programmatic imperative of UC3 is to provide a curation environment that is comprehensive in scope, yet flexible with regard to local policies and practices, responsive to requirements of funding agencies for data management and open access, and cognizant of the inevitability of disruptive changes in technology and user expectations. Harnessing the collective energy and innovation of its partners, UC3 provides solutions to the academic communities that are out of the reach of any individual partner.

Patricia Cruse is the founding director of the University of California Curation Center (UC3) and is responsible for all services within UC3. She works collaboratively with the ten UC campuses to develop sustainable strategies for the curation and preservation of digital content that supports the research, teaching, and learning mission of the University. Ms. Cruse has developed and oversees several of CDL's major initiatives, including the NDIIP-funded Web Archiving Service and the Digital Preservation Repository. Trisha serves on the HathiTrust Strategic Advisory Board. Her activities include specifying preservation services for the HathiTrust initiative and working with UC campus stakeholders to develop a set of digital curation micro-services supporting research data. Trisha's current work focuses on developing tools and services that support broad types of academic output. Finally Ms. Cruse is on the leadership team for the multi-institution, NSF-funded DataONE initiative.

II. NED GARDINER

The Future is Unwritten: Data and Information for a Transforming World

Generations of humans have demanded that the world wake up, that people stop participating in lifestyles that push us beyond planetary boundaries, and that decision makers come to their senses. We all know this strategy has failed to transform complex, coupled human-natural systems at a scale or rate that will slow the biodiversity crisis, reverse anthropogenic climate change, or return the chemical state of ocean basins to pre-industrial conditions. Any strategy predicated only upon providing information is likely to see similar results, yet good information is essential for human society to collectively explore options for addressing these complex issues while simultaneously providing for upwards of nine billion brothers and sisters on this small planet. This paradox is good news for information managers and the discipline as a whole. Semantic engines, Earth system grids, and other technologies aimed at retrieving and using interconnected assets can and do aid in complex information products designed for audiences around the world. Your skills are essential for the challenges of our age.

Ned Gardiner is the Visualization Manager for NOAA's (the National Oceanic and Atmospheric Administration) Climate Program Office and a producer of www.climate.gov, a flagship web site providing cutting-edge, accurate climate information. For a decade, he has used scientific visualization to help make complex scientific information understandable. Recently, he has focused on helping decision-makers around the country use climate data products make well-informed decisions about climate, climate change, and interactions with living systems. Earlier in his career, Ned advanced the use of satellite data and digital maps to produce biodiversity and Earth science video programming for museums around the world.

Environmental Information Management 2011

Sept 28-29, 2011
Santa Barbara, CA

Wednesday, September 28, 2011

7:00 AM	**Registration and Coffee**		(Location TBD)
8:15 AM	**Welcome to EIM**	M. Jones and C. Gries	

S1: Sensors and Workflows — Santa Ynez Room

8:30 AM	Grady et al.	Lifemapper, VisTrails and EML: Documented, Re-executable Species Distribution Models
8:50 AM	Barseghian et al.	Sensor lifecycle management using scientific workflows
9:10 AM	Barateiro et al.	Archiving Sensor Data - Applied to Dam Safety Information
9:30 AM	Jaroensutasinee et al.	Coral sensor network at Racha Island, Thailand
9:50 AM	Gries and Porter	Moving from Custom Scripts with Extensive Instructions to a Workflow System: Use of the Kepler Workflow Engine in Environmental Information Management
10:10 AM	Lerner et al.	Provenance and Quality Control in Sensor Networks

BoF1: Birds-of-A-Feather sessions — Fiesta Room

8:30-10:30 AM	Hollingsworth	Internet mapping: What are the options?

10:30 - 11:00 AM BREAK

Keynote: Patricia Cruse, California Digital Library — **Santa Ynez Room**

11:00AM-12:00 PM	Building Communities, Partnerships, Tools, and Services in Order to Thrive in a Dynamic Information Landscape

12:00 - 1:30 PM **LUNCH**

Wednesday (continued)

S2: Semantics and Data Management — Santa Ynez Room

1:30 PM	Gandara et al.	CI-Server Framework:Cyber-Infrastructure Over the Semantic Web
1:50 PM	Leinfelder et al.	Using Semantic Metadata for Discovery and Integration of Heterogeneous Ecological Data
2:10 PM	Zheng et al.	A Semantically-Enabled Provenance-Aware Water Quality Portal
2:30 PM	Simons et al.	Toward species interaction networks – Managing, visualizing and synthesizing Gulf of Mexico geo-spatial trophic data
2:50 PM	Lehtonen et al.	The Process of Digitising Natural History Collection Specimens at Digitarium
3:10 PM	Strasser et al.	DataONE: Promoting Data Stewardship Through Best Practices
3:30 PM	Tenopir et al.	Understanding the data management needs and data sharing challenges of environmental scientists

BoF2: Birds-of-A-Feather sessions — Fiesta Room

1:30 - 2:30 PM	Robinson	Using Web Tools and Methods to Support Earth Science Collaborations
2:30 - 3:30 PM	Valentine, Skibbe and Hollingsworth	Geospatial Data Management for Ecological Research Organizations

Poster Session and Reception — Sierra Madre North Room

4:00 - 7:00 PM Poster Session and Reception
-- Lightning talks start at 4:30 (1 minute each, no slides)

Thursday, September 29, 2011

8:00 - 8:30 AM **Coffee**

S3: Discovery, Visualization, and Analysis — Reagan Room

8:30 AM	Gallagher et al.	Searching for satellite data sets using Kepler, Metacat and EML
8:50 AM	Franklin et al.	Rapid Development of a Hybrid Web Application for Synthesis Science of Symbiodinium with Google Apps
9:10 AM	Belbin	The Atlas of Living Australia
9:30 AM	Stewart et al.	Lifemapper: Infrastructure and Services for Biodiversity Science
9:50 AM	Pham et al.	Interactive Visualization of Spatial and Temporal Patterns of Diversity and Abundance in Ecological Data
10:10 AM	Banerjee et al.	Development and application of a fast and accurate image analysis algorithm to study the influence of vermicompost and AMF on the growth of Azadirachta indica

BoF3: Birds-of-A-Feather session — Fiesta Room

8:30-10:30 AM	Sheldon and Porter	Automating Data Processing and Quality Control using Workflow Software: Converting Sensor Data to Usable Environmental Information

10:30 - 11:00 AM **BREAK**

Keynote: Ned Gardiner, National Oceanic and Atmospheric Administration — **Reagan Room**

11:00AM-12:00 PM — The Future is Unwritten: Data and Information for a Transforming World

12:00 - 1:30 PM **LUNCH**

Thursday (continued)

S4: CyberInfrastructure Systems — Reagan Room

1:30 PM	Bainbridge	A services based architecture for delivering interoperability for environmental observational data
1:50 PM	Glaves and Schaap	Geo-Seas: a pan-European marine geoscientific e-infrastructure
2:10 PM	Izurieta et al.	A Cyber-Infrastructure for a Virtual Observatory and Ecological Informatics System -VOEIS
2:30 PM	Porter and Kortz	Web Services in the U.S. Long-Term Ecological Research Network: Now and in the Future
2:50 PM	Tarboton et al.	Data Interoperability in the Hydrologic Sciences: The CUAHSI Hydrologic Information System
3:10 PM	Zaslavsky et al.	The Initial Design of Data Sharing Infrastructure for the Critical Zone Observatory

BoF4: Birds-of-A-Feather sessions — Fiesta Room

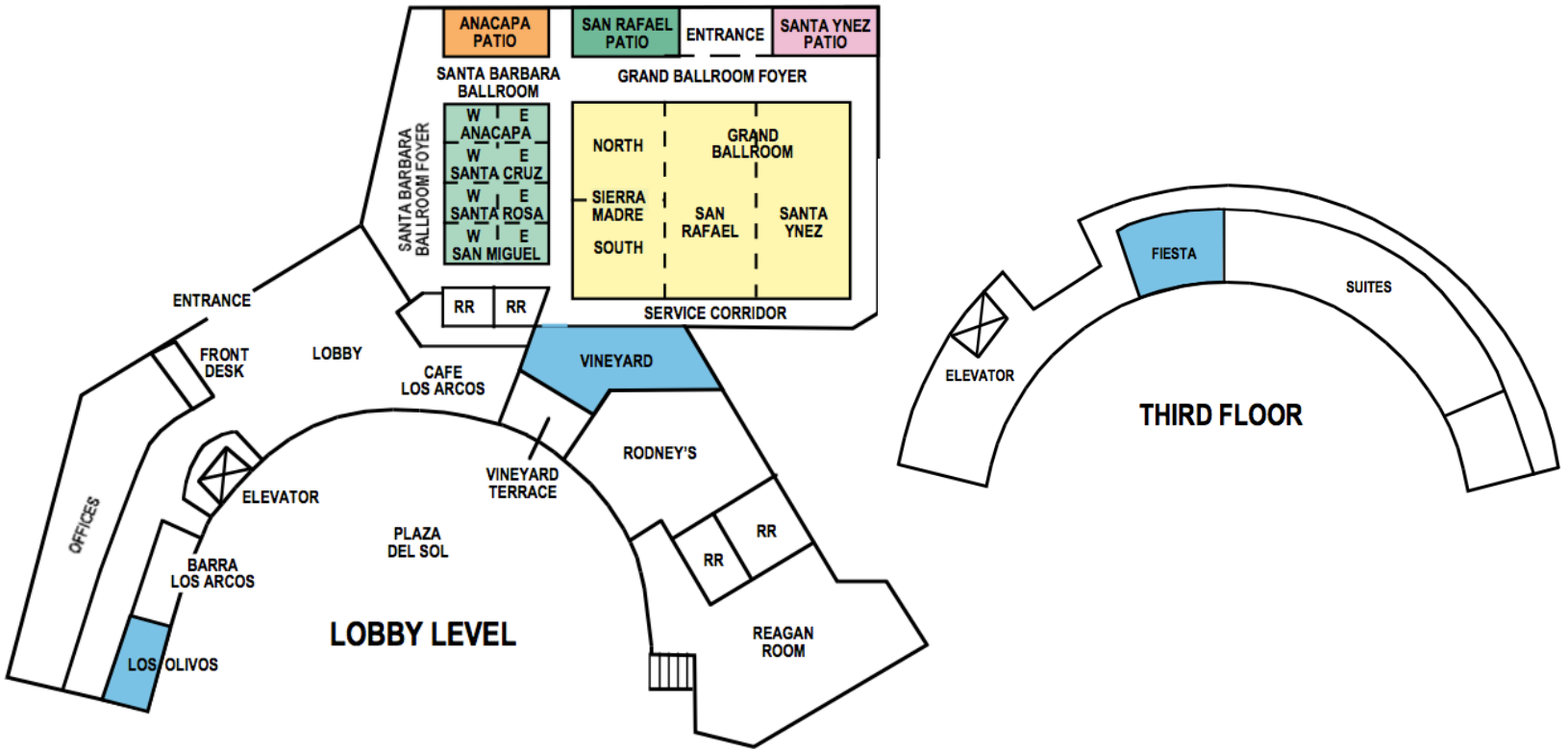
1:30 - 3:30 PM	O'Brien and Servilla	Functional Requirements for the EML Dataset Congruency Checker

3:30 - 4:00 PM **BREAK**

Plenary Discussion — Moderator: O'Brien — Reagan Room

4:00 PM - 5:00 PM	O'Brien	Community Standards and Practices Development
5:00 PM	**Concluding remarks**	Jones and Gries

Conference Room Layout
Fess Parker Doubletree Hotel



Sponsored by: NCEAS LTER DataONE

PREFACE

The Environmental Information Management Conference 2011 provided a cross-disciplinary forum for information managers, computing researchers, software developers, and environmental scientists interested in technologies that enable data collection, description, curation, discovery, access, integration and analysis in all disciplines of environmental research. Participants from throughout the world convened in Santa Barbara, California to showcase advances that cross computing, environmental science, and informatics disciplines. In addition to presenting new work in environmental informatics, EIM provided a forum to build partnerships, explore solutions to the common challenges faced by environmental observatories, and to present advances in community standards, practical system design, implementation and assessment.

This proceedings volume contains 25 contributed papers and the abstracts of 20 contributed posters that together were the core of the informatics advances presented at the conference. Papers were rigorously peer reviewed by the EIM Program Committee, who we thank for their prodigious time investment, which culminated in the high quality papers presented this year.

As we embarked on the 2nd Environmental Information Management this year, we were struck by the enthusiasm to continue this tradition of an applied informatics conference that highlights new approaches to computing in environmental science. As the emphasis on open science and open data continues to grow, there is a growing need for the cross-disciplinary forum represented by EIM. While we wish that these proceedings are useful today to environmental sciences, we are already thinking about the next incarnation of the conference, and we invite you to participate in that discussion by contacting us with your ideas, concerns, and inspiration.

Matthew B. Jones
Corinna Gries
EIM 2011 Co-chairs

CONTRIBUTED PAPERS

A services based architecture for acheiving interoperability of environmental observational data

Scott Bainbridge
Australian Institute of Marine Science, PMB 3 MC, Townsville 4810 Australia
s.bainbridge@aims.gov.au

Abstract— **Observational data typically conform to a simple pattern of location, time and observed value. This allows for various types of environmental observational data to be held within a single data framework. If a services based access protocol is wrapped around this framework it becomes possible to build a simple data management system that implicitly allows for and promotes data integration and interoperability. This paper describes such an architecture with an example of how this is being developed to achieve interoperability between coral reef sensor network data from four global sites.**

Keywords—sensor networks; coral reefs; web services; OGC SWE

I. INTRODUCTION

Advances in automated sampling, sensor networks and remote instruments make it possible to collect large volumes of environmental observation data, much in real time, to the point where now the issue is too much raw data and not enough derived information. Observation data, especially in real time, has a key role in providing information to help deal with a range of environmental and societal issues. In order to respond to events such as cyclones, floods and disasters, such as the Japanese tsunami, authorities need reliable real time data and information products to optimize responses and even save lives. In one example real time radiation maps around the damaged Fukashima Daiichi power plant were produced using a combination of formal data sources and crowd-sourced Geiger counters [1,2]. The use of crowd-sourced data, while unreliable, was for a period the only data publically available [2] giving important information about the incident.

The development of such maps, using a range of data sources, requires, and indeed mandates, full data integration. Initial attempts at data integration revolved around setting rigorous end-to-end standards, not only in how the data were collected, but how they were stored (schemas) and processed. Many of these attempts failed as there is no 'one size fits all' solution and many organizations are limited in what technologies and approaches they can utilize. It therefore became impossible to simply 'impose' prescriptive external standards as a way of forcing data integration.

The delivery of data as web or *'HTTP'* based services promised to solve some of these issues. By 'wrapping' internal systems with standards based services it becomes possible to abstract the internal systems and processes from the external interfaces. This allows systems to 'talk' to each other even if they utilize differing internal technologies. This approach has been used extensively in the business world to implement business to business (B2B) solutions that allow interoperability between often disparate systems via standardized interfaces and protocols [3]. In the geospatial world, standards such as Web Map Service (WMS) [4] and Web Feature Service (WFS) [5], have been used to achieve the same result [6].

The issue with WMS or WFS is that they are mapping standards and so explicitly spatial but only implicitly temporal; most observational time series data are the opposite. Many observation time series datasets have thousands of observations at a single spatial point while most maps have many spatial points at a single point in time. For this reason mapping standards do not suit time series observational data [7].

To resolve these issues the Open Geospatial Consortium (OGC) developed the Sensor Web Enablement (SWE) series of protocols [8] for time dependant observational data including environmental observations. The SWE 'stack' includes functionality such as event detection, full temporal and spatial querying, the delivery of data as data blocks rather than as spatial objects, descriptions of the sensors and observation process via SensorML records, along with standards for re-tasking and controlling sensors [9].

The SWE therefore provides a set of standards and protocols on which a service based data management system suitable for a range of environmental observation data could be built. The adoption of a common set of protocols for information exchange, via the SWE standards, should facilitate the drive towards better data integration and use. This paper looks at work to integrate environmental data from a series of coral reef sensors networks using a cloud based, service orientated data management system.

II. CORAL REEF SENSOR NETWORKS

The Coral Reef Ecological Observatory Network (CREON) is a community group facilitating the deployment of coral reef sensor networks. It acts to coordinate existing work being done at Moorea in French Polynesia through the US Long Term Ecological Research (LTER) network, at Kenting National Park in southern Taiwan through *Academia Sinica* Taiwan and the Taiwan National Centre for High-Performance Computing (NCHC), in Thailand at Racha Island near Phuket through the University of Walailak and the Thailand National Science and Technology Development Agency (NSTDA), and on the Great Barrier Reef in Australia as part of the Australian Integrated Marine Observing System (IMOS) through the Australian Institute of Marine Science (AIMS).

One of the CREON goals is to integrate data from each of the sites to look at global issues such as coral bleaching, impact of climatic events, and so on. In 2010 sensors were set up at each site to measure basic environmental parameters, such as in-water temperature, salinity and pressure and above water air temperature, humidity, rainfall, wind direction and speed as well as light as Photosynthetically Active Radiation (PAR).

However, given the broad range of organizations involved, it was not possible to achieve data integration through imposing internal organizational standards such as common data schemas. The next approach examined was that of a federated data model with each agency managing their own data and with data integration being achieved by exposing the internal data via standardized interfaces (such as via the SWE protocols and standards). This model allows agencies to keep their internal systems with integration occurring at the external interface level. The main issue with this model was that this still requires agencies to maintain the servers that implement the external interfaces and many agencies had security concerns such as allowing access to externally-facing servers, resulting firewall issues, and so on.

The final model examined was to 'push' the data from each agency to a single cloud-based services-orientated data management system. This proved to be easier in terms of security as most security systems are design to prevent intrusion, not control outgoing data. As the data itself are not restricted or sensitive the Institutional security need is not to protect the data but rather to protect the data infrastructure, such as the internal networks and servers. The use of an external cloud data store effectively transfers the security issue to the provider and so, for non-sensitive data, the push model resolves some of the organizational security issues.

The problem was that there are no publically available environments suitable for hosting the CREON data. Work was undertaken to investigate what functionality such a system would need to provide and how it could be constructed given the availability of existing software to implement the SWE standards. As a result the CREON group looked at how cloud computing services based architectures could deliver on the need to integrate data from each of the sites into a single system that could then deliver information about global scale processes impacting coral reefs.

III. System Design

A. Needs Analysis

The first step was to look at the required functionality, referred to as a needs analysis. The needs analysis was done as a three stage process. The first stage was to look at the overall functionality that the system would need to provide to meet some basic use-cases. The second stage was to drill down and look at the data and functional entities that would be needed to meet the required functionality. This step involved comparing what the SWE components could deliver against what the use-case indicated was required. The final stage was to look at how the system may be utilized in the future and in particular how the system might work within emerging paradigms such as the Internet of Things [10].

The base use-case was developed around the need for a person to simply find a dataset using a natural language query, to be able to quickly display the data and assess the fitness for purpose using the displayed data and ancillary information, such as how the data was collected, who collected it and so on. This was condensed down to the idea of data **d**iscovery, **d**isplay, **d**ownload and **e**xploration; or what was called **D^{3}E**.

Using the initial needs analysis the following user level functionality was described:

- Ability to **discover** data using natural language simple search interfaces, typically what, where, how – so what data exists for this area/time/theme, who collected it and how, what 'quality' does the data set have and how can I use it;
- Quickly plot up or **display** data to see any overall patterns in the data, to assess how suitable it is for the required purpose, how it relates to other datasets and what quality control has been applied to the data (how 'fit for purpose' it is);
- **Download** the data either as raw data, as a processed product (such as re-sampled to a set space/time grid, time or space averaged, with certain quality control rules applied, etc) in a set of standard file formats (e.g. comma separated, spreadsheet, etc);
- Perform basic analysis or data **exploration** such as comparing two time series, time shifting data, re-gridding data (space / time), plotting regressions, etc;
- To define and register events of interest from the discovered data streams, and then define actions based on the event triggers, these maybe simple notification actions or more complex service chaining actions;
- Get the complete set of ancillary data such as a full SensorML [11] record, an International Standards Organization (ISO) 19115 metadata record [12] and potentially other data such as calibration data, deployment data including photographs and so on.

The user level functionality was then translated into system level functionality, that is things the software system needed to do to provide the required functionality back to the user. The following system functionality was identified:

- Register new data streams, preferably via a services based interface;
- Upload and store deployment and other details of the data stream so that a valid SensorML record can be generated;
- Upload and have available as a service a full metadata record linked into the deployment and data stream with a preference for an ISO-19115 [12] compatible record;
- Upload and store the sensor data itself, preferably directly from the sensor platform itself not via a central data centre, that is directly from the remote field instruments;

- Perform quality control over the stored data using serviced based agents;
- Produce simple graphical outputs, such as time series graphs, from the data, again as a service;
- Deliver the data, as an eXtensible Markup Language (XML) file, based on a range of queries including spatial, temporal and thematic queries;
- Synchronize the cloud based data with Institutional data stores so that copies of the data can be stored in traditional data systems for backup and archiving;
- Register events and get notifications if an event occurs, at a higher level the ability to link event triggers into other notification and decision support systems;
- Perform simple statistics on multiple data streams to allow for basic synthesis and information extraction.

The service nature of the architecture gives it flexibility to adapt to new standards and even new paradigms such as the Internet of Things [10,22]. New standards can be supported by writing new wrapper services around the data stores, similarly new types of requests can be dealt with by adapting the service request layer. As long as the fundamental data units and processes are incorporated into the system, along with the correct linkages, a service based architecture should able to adapt to future needs.

B. Identifying the main components of the system

From the needs analysis the next step was to identify the main functional units or components that would logically deliver the functionality derived from the needs analysis and then see how these map onto existing solutions. Fig 1 shows the main envisaged components. The **Services Request Layer** would identify the type of request and direct the request to the appropriate component, a **Security** layer sits between the system and the external world to ensure that only appropriate requests are processed.

The **Ingest** component would listen for data and on a valid request insert the data into the data store; it may or may not do some checking before this operation. The **Registry** component would respond to queries about data sets held by the system and return a list of matching entries. The **Scheduler / Workflow** component would organize other components based on time or events generated by other components (such as the event detection / quality control component). The **Quality Control** component would run the quality control routines on a regular basis or as required. The **Graphical Display** component would produce graphs and other display 'widgets'.

The **Data Download** component would produce standard format files for downloading from the data while the **External Sync** component syncs the cloud based data-store to data systems residing in other organizations. The **Statistics / Processing** component would be a general purpose component to allow for standard and custom processing of the data. It could be based on Kepler [13] or similar system where the user can define a workflow and access statistical routines. Some of these may be pre-defined but others may be user defined.

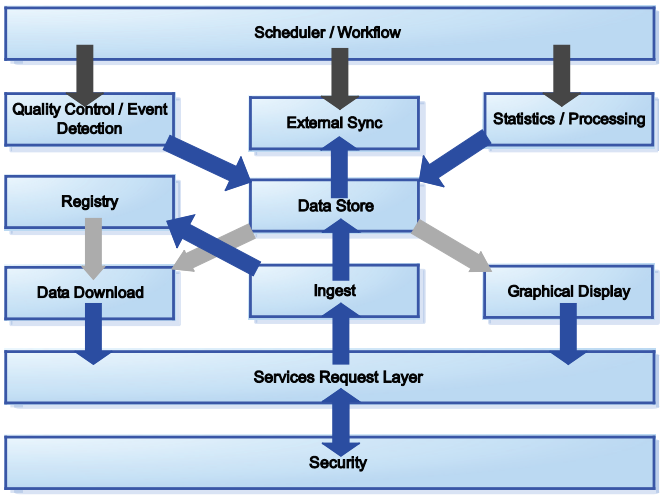


Figure 1. Design of the main components of the system, arrows show the main data flows.

IV. REGISTRIES

One of the fundamental requirements from the use-case was the ability to do simple natural language searches to discover data. To implement this, the design (Fig. 1) has a Registry that is used to store a series of thematic, temporal and spatial search terms along with linkages between the data stream, metadata record, and SensorML record.

The term 'registry' in web data systems often refers to a list of available (web) services. In this case the term is used to represent an entity that sits between the user and the metadata catalogue and other supporting data, to deliver a set of data records that reflect a query passed to it. So the registry is almost a 'super' or higher order metadata catalogue that is optimized against a number of use-cases (each of which may have its own registry instance) to deliver targeted responses against that use-case query.

There is seemingly a degree of overlap in the functionality provided by the SensorML record, the metadata record and what a registry would do. The SensorML record describes the sensor and, as importantly, the way that the original electrical measurement (normally a voltage measurement) gets processed into the final real world value including any changes of units, conversions using calibration coefficients, and so on. As such it provides a link between the basic measurement event and the final observation value.

The metadata record provides general information about the data collection event such as the organization or person that collected the measurement, textual summaries of the sensor deployment, citation details, use constraints, descriptive keywords, the location and format of the data if available and so on. The ISO 19115 format [12] has many of these fields as unstructured text fields and so it is not possible to know in advance the format of the contents, this makes it difficult to incorporate into automated searches.

So why have a registry? The original use-case involved the use of natural language queries that return a limited of 'hits' that the user could then select from. With this was the ability to assess the 'fitness for purpose' of the data returned. One area not described, but of potential benefit, is the ability to link to ontologies so that user can use vocabularies and terms from other domains to search for data. This level of functionality cannot be achieved using the SensorML record and the ISO 19115 metadata record alone and so a registry, as a separate entity, is required to fill this need.

Registries have proven to be difficult to build and tend to be complex. There are currently a number of approaches being taken with the development of sensor data registries. The first is mapping the SensorML record to an ebRIM catalogue [14] to allow for searching using the SensorML attributes. The second is the development of two new draft standards within the SWE stack; the Sensor Instance Registry (SIR) [15] and Sensor Observable Registry (SOR) [16] standards. Both of these approaches do not currently have implementable software although an incubator project is underway [17] using the SIR and SOR standards.

Projects such as the Oceans Tethys project [18] have used a harvested getCapabilities document from registered SOS servers to build a basic registry and to allow for simple searches. For the CREON project the concept of a registry was that of a simple set of thematic search terms along with linkages between the metadata record, the SensorML record and the data stream itself. This would allow the system to find datasets based on common search terms and then link the data, the metadata and the SensorML information together.

V. BUILDING A SYSTEM

The first task was to try and map the functional units and design elements to existing standards and software. The initial target was the OGC SWE set of standards, along with the ISO 19115 metadata standard, as these seemed to be the most advanced and commonly used standards for observational data.

From Fig. 1 the **Ingest** and **Data Download** units mapped across to the SWE Sensor Observation Service (SOS) software, the **Registry** component could be a combination of the SOS getCapabilities function, the SensorML record and the metadata record. It is possible to deliver the metadata record as a service using the GeoNetwork software [19] with the appropriate metadata profile but as this sits outside the SWE stack it has to be somehow linked back into the SWE data. The **Quality Control / Event Detection** could be executed via the SWE Sensor Alert Service (SAS) or Web Notification Service (WNS) and this could also be used to implement a crude **Scheduler / Workflow**.

The SWE stack only covers standards and protocols for data interchange; it does not explicitly include standards for any graphical components or any other service based data delivery. To implement these requires separate software, either as a client or via services, that take the SOS data and format these into other products. While the SWE standards do include standards for workflows and sensor processing these are limited to particular use-cases, such as tasking sensors, rather than general workflows such as quality control.

The mapping shows that it is possible to deliver a cloud computing based services-orientated system for storing and delivering environmental observation data using the Sensor Observation Service of the SWE stack. What is less clear is how to deliver graphical products, deploy a functional registry, do complex quality control and other processing, implement workflows and put in place some level of security.

At this point the most obvious way forward was for one of the agencies to install an externally facing SOS server (such as the 52° North implementation [23]) to store and provide access to the observational data. The next step would be to write small programs for each of the agencies to 'push' their data to the central server. Then a web based server system could be built to allow for the **D^{3}E** functionality including provision of graphical content, registry functionality and data download. This effectively meant writing significant amounts of custom code to interact with SWE compliant data stores to deliver the total suit of required functionality.

While there are semi-mature implementations of the base software (such as the SOS server) and a few basic SOS clients the initial vision was not something that could be easily implemented by the group. The reality was that project did not have the resources, or the mandate, to develop extensive software so a range of alternatives were investigated. One alternative that is currently being trialed is the Pachube (pronounced 'patch-bay') [20] system that seems to have much of the required functionality.

The Pachube system allows for data streams to be registered, for data to be stored and retrieved and for simple graphics products to be delivered, such as time series graphs. It implements some basic security using project level keys, includes event triggers that can activate other processes via service chaining, and allows for simple registry to be created using machine tags [25]. This combined functionality, hosted by a commercial entity, was investigated as a possible way to deliver on the initial promise of delivering integrated data from each of the CREON sites.

The issue with the Pachube system is that it does not currently support the SWE set of standards but rather uses EEML [21] for data upload / download. This may make it harder to integrate into SWE compatible tools, services and systems to the point where, as SWE becomes better supported, interoperability with Pachube may become an issue. Pachube is also a commercial service although charges are nominal at the moment. On the positive side the site is fully functional now, is heavily integrated into the Internet of Things (IoT) [22], and is looking to actively develop solutions for particular application areas.

In trying to build an operational system it became evident that although many of the required standards and functionality do exist these are still a long way from delivering simple 'Lego®-block' tools and software to build and deliver systems for environmental observational data. Having said this there are no technical reasons why it can't be done, it is just that the existing approaches are yet to deliver a 'pre-fabricated' solution. This will happen but there is still a need to drive this process and the value of projects such as CREON maybe in doing just this.

VI. THE CREON APPROACH

The need to deliver integrated interoperable data for the CREON community group lead to two possible approaches; either build considerable software to implement a standards based SWE compliant system, or look at commercial systems that may provide equivalent functionality, even if not SWE compliant.

The current approach is to use the Pachube system, data from two sites (Racha Island in Thailand and the Great Barrier Reef in Australia) are currently being fed into this system. A simple web display of the data is available via the CREON web site (http://www.coralreefeon.org) allowing data from the two sites to be plotted side by side. While the Pachube system fills a number of immediate needs the group is still looking for opportunities to deliver or deploy a more standards based solution. The group is still focused on the SWE set of standards but in order to deliver a solution now it has utilized the commercial Pachube solution.

VII. THE 'SO-WHAT' QUESTION

It may seem odd that a series of coral reef biologists end up investigating services based approaches to the delivery of data. The primary reason for CREON is that there is no point collecting data if no-one can find it or use it. This is the fundamental 'So-What' proposition. All of the millions of dollars spent on data collection and management are somewhat in vain if the data cannot be found, explored and used.

The first issue is data discovery. The standard response is that metadata catalogues fill this need; they do and they don't. The main issue is that metadata is collected in response to a range of needs from fulfilling user queries to low level machine to machine transactions. Metadata standards, such as ISO 19115, rely on profiles or community driven instances to define the actual content in each element and so a system accessing ISO 19115 may have to do complex decoding against the profile, presuming one exists. The metadata does not explicitly include links to a SensorML record or other ancillary information, although again this can be done through the use of profiles. Finally metadata is collected at a range of levels or granularity which can vary from project to project.

The need for metadata to be all things to all people means that using it to deliver specific responses against set use-cases is difficult, especially when dealing with a range of metadata standards and profiles. The solution is to insert another entity, the Registry described here, between the metadata catalogue, the SensorML record and other ancillary data, and the user. This then allows specific use-cases to be implemented ensuring that standard responses are delivered against queries. This paper argues that registries are an area that requires more work and initiatives such as the Sensor Instance Registry [15] and Sensor Observable Registry [16] may help.

Not only do we need better tools to find the data but also better tools to explore and extract the knowledge from the data. As we become better at collecting data we need to also become better at analyzing and delivering the data, or better, the information and knowledge within the data. This knowledge will more and more come from multi-parameter data sets collected by a range of agencies many of whom may not scientific institutions but rather, as with the example of the Japanese radiation maps, any person with the capacity and interest to contribute. This dramatically changes the nature of the data (such as the quality of the data), how it should be used, how it can be delivered and how it can be integrated into other data sets. There are both threats (poor quality data, data used inappropriately) and opportunities (more data, cheaper data collection, ad-hoc data collection) in this model.

The answer to the 'So What?' question is that new advances and understandings will be made from integrating data sets in new ways, often as totally new data products, to deliver new understandings and knowledge. To achieve this requires a move from institutional data centers to open inclusive service based data systems. The relatively simplicity of environmental observational data (numbers over images and video) presents an opportunity to lead this process.

VIII. THE 'INTERNET OF THINGS' - IOT

Currently systems are being built for humans to access but the next generation of systems will be built for machine to machine interaction, the so called 'Internet of Things' [10]. As an example it is now possible to check the coming weather using a number of services or 'apps' via smart phones, tablets, as well as traditional computers. A check in the morning can tell you if you need to take an umbrella or not. In the Internet of Things paradigm the umbrella will request the daily weather each morning and if, through a machine to machine interface with the weather service, it detects that rain is likely, it will notify its owner that it maybe wet and that they should take the umbrella. The interaction is no longer between the person and the weather service, but rather a machine to machine interaction between the umbrella and the weather service.

In an observing context event detection systems will monitor a series of data streams looking for events of interest, if an event fires then this can be linked into modeling and scenario systems to predict the outcome of the event and then linked further into appropriate responses. The recent events in Japan show how such a system could operate and the role that real time data, sitting as input to decision support systems, could have in such a situation.

The vision is that if all of this data exists as publically available serviced orientated data streams, complete with quality control, metadata and ancillary data, then it becomes possible to build a totally new set of knowledge tools. These tools could deliver totally new outcomes from our data and it is this that drives groups such as CREON.

IX. CONCLUSION

The idea that agencies will continue to fund environmental observational programs that do not make their data publically available is outdated. As issues such as climate change, ecosystem sustainability, and environmental impacts gain prominence, alongside events such as the tsunami in Japan and the Gulf of Mexico oil spill, the demand will be for multi-disciplinary datasets that are freely available via service based interfaces linked into decision support and modeling systems.

A simple need to integrate coral reef environmental observatory data from a small number of sites has lead to the larger vision of making large amounts of environmental observation data available as centralized services. The software to do this is either available or being developed (for example the open source 52° North initiative [23]); what is lacking is combining this with cloud based data storage to offer a complete solution as detailed here. Issues such as security and cost are impediments but the Pachube example shows that these can be overcome.

There is an opportunity within the environmental observatory community at large (for example the marine observing community as represented by the OceansObs'09 conference [24]) to take leadership and to develop an open, standards-based, cloud-computing data management system to deliver a range of environmental, and other, observational data. This data becomes a resource that contributes to our understanding of how environmental systems are changing and to help in their protection, conservation and long term sustainability.

ACKNOWLEDGMENT

The sensor network component of the Integrated Marine Observing System (IMOS) is funded by the Australian Government through the National Collaborative Research Infrastructure Strategy and by the Queensland State Government. The Racha Island work is funded through the University of Walaliak, Thailand, and the Thailand National Electronics and Computer Technology Center (NECTEC). The work at Kenting National Park, Taiwan, is funded by the National Centre for High Performance Computing (NCHC) and *Academia Sinica* Taiwan with the work at Moorea is funded by the US LTER Project via the University of California – Santa Barbara.

REFERENCES

[1] H. Zhang. "Japan Geigermap," http://japan.failedrobot.com, accessed July 2011.

[2] CNET, "Japan monitoring goes crowd, open source," http://news.cnet.com/japan-radiation-monitoring-goes-crowd-open-source/8301-17938_105-20060639-1.html, accessed July 2011.

[3] C. Bussler, "The Role of B2B Protocols in Inter-Enterprise Process Execution," in Lecture Notes in Computer Science: Technologies for E-Services, F. Casati, M.C. Shan and D. Georgakopoulos, Eds. Springer, Berlin 2001, pp. 16-29.

[4] J. de la Beaujardiere, "OpenGIS Web Map Server Implimentation Specification," Open Geospatial Consortium Document 06-042 Version 1.3.0, March 2006, pp. 1-85.

[5] P.A. Vretanos, "Web Feature Service Implimentation Specification," Open Geospatial Consortium Document 04-094 Version 1.1.0, May 2005, pp. 1-131.

[6] T.R. Duffy, E. Boisvert, S. Cox, B.R. Johnson, O. Raymond, S.M. Richard, F. Robida, J.J. Serrano, B. Simons and L.K. Stolen, "The IUGS-CGI international geoscience information interoperability testbed," International Association for Mathematical Geology, XIth International Congress, Université de Liège – Belgium, September 2006, S05-06 pp. 1-4.

[7] L. Bermudez, T. Cook, D. Forrest, P. Bogden, C. Galvarino, E. Bridger, G. Creager and J. Graybeal, "Web feature service (WFS) and sensor observation service (SOS) comparison to publish time series data," International Symposium on Collaborative Technologoies and Systems (CTS) '09, Baltimore USA, May 2009, pp. 36-43.

[8] M. Botts, G. Percivall, C. Reed and J. Davidson, "OGC Sensor Web Enablement: Overview and High Level Archetecture," OGC White Paper: OGC 07-165, Open Geospatial Consortium, 2007.

[9] Open Geospatial Consortium, "OGC Sensor Planning Service Implimentation Standard," http://www.opengeospatial.org/standards/sps, accessed May 2011.

[10] N. Gershenfeld, R. Krikorian and D. Cohen, "The Internet of Things," Scientific American 291:44, 76-81, October 2004.

[11] M. Botts, "OpenGIS Sensor Model Language (SensorML) Implimentation Specification," OGC 07-000, Open Geospatial Consortium, 2007.

[12] International Standards Organization, "ISO 19115:2003 Geographic information – Metadata," ISO 19115:2003(E), 2003.

[13] Kepler Project, "Real-time Environment for Analytical Processing (REAP)," https://kepler-project.org/users/projects-using-kepler-1/reap-project, accessed May 2011.

[14] F. Houbie, F. Skivee and S. Jirka, "OGC Catalogue Services Specification 2.0 Extension Package for ebRIM Application Profile : SensorML," Discussion Paper OGC 09-163r2, Open Geospatial Consortium, 2010.

[15] S. Jirka and D. Nüst, "OGC Sensor Instance Registry Discussion Paper," Discussion Paper OGC 10-17, Open Geospatial Consortium, 2010.

[16] S. Jirka, A. Bröring and D. Nüst, "OGC Sensor Observable Registry (SOR) Discussion Paper," OGC Discussion Paper 09-112r1, Open Geospatial Consortium, 2010.

[17] 52 Degrees North, "Sensor Discovery Service Intergation and Workflow", http://52north.org/communities/sensorweb/incubation/discovery/index.html, accessed May 2011.

[18] OOSTethys, "OOSTethys and the OGC Oceans Interoperability Experiment – OOSTethys Architecture," http://www.oostethys.org/System%20Architecture, accessed May 2011.

[19] GeoNetwork opensource, "GeoNetwork opensource," http://geonetwork-opensource.org/, acessed May 2011.

[20] Pachube, "Pachube – data infrastructure for the Internert of Things," http://www.pachube.com/, accessed May 2011.

[21] EEML, "Extended Environments Markup Language – EEML," http://www.eeml.org/, accessed May 2011.

[22] Read Write Web, "Top 5 Web Trends of 2009: Internet of Things," http://www.readwriteweb.com/archives/top_5_web_trends_of_2009_internet_of_things.php, accessed May 2011.

[23] 52° North, "52° North Initiative for Geospatial Open Source Software," http://52north.org, accessed July 2011.

[24] OceanObs'09 "Ocean information for society: sustaining the benefits, realising the potential. 21-25 September 2009, Vencie, Italy.," www.oceanobs09.net, accessed July 2011.

[25] Wikipedia, "Tag – metadata," http://en.wikipedia.org/wiki/Tag_(metadata), accessed July 2011.

Development and application of a fast and accurate image analysis algorithm to study the influence of vermicompost and AMF on the growth of *Azadirachta indica*

K. Banerjee[1], Y.T. Jasrai[2] and N.K. Jain[3]
[1] Gujarat Forest Research Institute, Research Division, 'j' Road, Sector 30, Gandhinagar 382 020, Gujarat, India
[2] Dept. of Botany, University School of Sciences, Gujarat University, Ahmedabad 380 001, Gujarat, India
[3] Dept. of Life Sciences, University School of Sciences, Gujarat University, Ahmedabad 380 001, Gujarat, India
kasturi_282004@yahoo.co.in, yjasrai@yahoo.com, drnkj11@gmail.com

***Abstract*—Measurement of root length and leaf area in plants often proves to be tedious and highly error prone when done manually with conventional methods. An inexpensive and accurate root length and leaf area measurement technique was developed in this work using digital image processing. The scripts were written in MATLAB and provide the ease of user specific extensions or modifications. Performance of three 'length estimators' were considered and the best suitable one was selected for our use in the nursery experiment with *Azadirachta indica* A. Juss seedlings. Accuracy within 5% was achieved and hence serves the purpose for this present experiment. The image processing protocol established here was extremely fast and provides the ease of analyzing a large batch of images. Further the leaf area measurement also yielded accuracy of 2-8% of the actual area. Sources of errors and their mitigation were discussed. Finally, the technique was applied to *Azadirachta indica* seedlings and variation of the parameters with age and treatments is reported.**

Keywords*— *image processing, root length, leaf area

I. INTRODUCTION

Root and leaf system parameters are apparent indicators of the plant growth performance, photosynthetic capabilities and nutrient and water uptake efficiencies. Measurement of root length and leaf area in plants often proves to be tedious and highly error prone when done manually with conventional methods (such as line intersect method, inch counter method etc. for root length and geometric methods for leaf area measurement). Most of the root length measurement procedures are based on the line intersect principle [1,2]. Manual methods involve reasonable human intervention and hence, prove to be error prone and highly time consuming, making it difficult to compare root lengths of the same species when determined in separate laboratories or when different measurement setups are used [3].

Several methods have also been proposed automating the line intersect principle, such as, the mechanical device based on a photo-diode equipped opto-electronic scanner [4], computerized image analysis for a video camera system [5-7], light sensor equipped X-Y plotter and slide projector [8] and desktop or handheld scanner system with a microcomputer [9,10]. These methods have expedited the root length measurement as compared to the manual measurements but rely upon the random placement of root samples ensuring least overlap, hence, involve a time-consuming sample preparation protocol. These methods are now being either replaced or realized in a modified way with the digital image analysis techniques, owing to the rapid development in computer and digital imaging hardware and software. These essentially involve maneuvering intensities recoded in pixel elements of a digital image and have proved to be reasonably accurate and faster. Several image analysis algorithms have been developed, each having its own advantages and inherent weaknesses.

Sophisticated commercial image analysis software like WinRHIZO (Regent Instruments, Quebec, Canada) and Delta-T SCAN (Delta-T Devices Ltd., UK) has been developed to analyze main root parameters like length, diameter, surface area etc. Some of it does not allow user defined extensions or modifications of the source file. Open architecture software like the NIH-Image [11] and ImageJ allows inclusion of user and task specific routines as well [12-14]. For our *Azadirachta indica* nursery experiment, however, a relatively fast, accurate and yet simple method for measuring the total root length and root surface area was required.

Here a comparative evaluation of several length estimators has been carried out to choose the most efficient digital image analysis algorithm for the root length of *Azadirachta indica* seedlings from a nursery experiment. Inherent limitations of such an algorithm, like the issues involved in skeletonization of binary images and overlaps have been addressed. A code is

developed in MATLAB accordingly and applied for the measurement of root length in the nursery plants.

A second part of the study comprises of the measurement of total leaf area of plant samples. Leaf area being an important agronomical parameter is directly related to plant growth, photosynthetic capacity and bio-productivity [15-17]. It is often used to assess the effect of different treatments on plant samples. Recently, hand held scanners and laser aided optical instruments are used for leaf area measurements. These instruments are generally expensive and rather complex in operation. A digital image based measurement of leaf area saves time compared to geometric measurements and increases accuracy at a nominal cost [15,18,19]. Here we demonstrate a technique involving scanned images [20] of all the leaves taken together and further subjected to MATLAB based image analysis routines. Both root length and leaf area are measured from their respective images within 8 seconds on a laptop running with Intel Core™ i5 processor 2.40 GHz with 4GB RAM and operating system *Windows 7 Home Basic* 64 bit.

II. THEORY

Several methods and length estimators have been proposed in the literature to date. To assess the applicability and reliability of the popular methods, a straight line AB is projected on a square grid of grid unit (equivalent to pixel length) p_d. The pixels containing information of the straight line are represented by a set of eight-connected pixels [21]. Each pixel in its neighbourhood is classified as orthogonally connected or diagonally connected with respect to its neighbouring pixels (figure 1a). These connections are categorized as either orthogonal connections (N_{oc}) or diagonal connections (N_{dc}) [13].

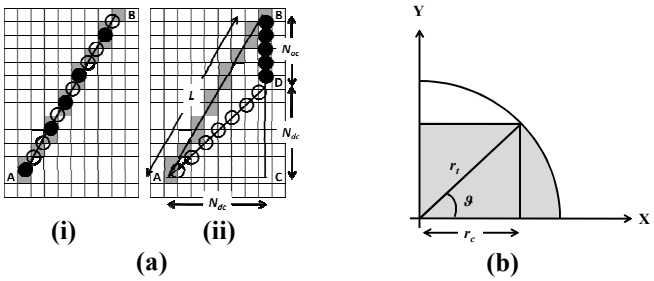


Figure 1. (a) Length calculation for a linear segment. (i) The segment AB represented on a mesh grid of square pixels. ○ stands for a diagonally connected pair and ● stands for a orthogonally connected pair of pixels. (ii) Hypotenuse of the right angled triangle ABC is represented by the line segment AB. Orthogonal and diagonal connections are arranged such that the side AC is represented by Ndc and the side BC is represented by the sum of Ndc and Noc. (b) Formulation of the correction factor for root length measurement. rt is the true root length; rc is the calculated root length and ϑ is the angle of the root segment with the x-axis.

The easiest approach would be to count up all the pixels (N_{tot}) that trace the line in a skeletonized binary image. Then simply multiplying the pixel count with p_d should provide the length of the line. This method yields a precise root length only when the root strands are aligned either vertically (along a single column) or horizontally (along a single row) in the image. In all other cases, for instance, when randomly scattered roots are measured, errors are introduced [22,23], i.e. calculated length (r_c) is always smaller than the true length (r_t), and thus a correction factor is required to be introduced [12,23]. Assuming that the root segments, i.e. the single pixel lines, are evenly aligned in all directions, increase of ϑ from 0 to $\pi/4$ (covering first quadrant) corresponds to the increase of y from 0 to $r_t \sin(\pi/4)$, where ϑ is the angle between the horizontal (x) axis and a root segment (refer to figure 1b). Area of the shaded part can then be represented as:

$$\int_0^{r_t \sin(\pi/4)} \sqrt{r_t^2 - y^2}\, dy \tag{1}$$

Here, $\sqrt{r_t^2 - y^2}$ is the length of the projection (i.e. r_c) of r_t on the x-axis for a given ϑ. Then, mean of r_c, calculated root length for the ones aligned from 0 to $\pi/4$, can be obtained by dividing the area by the height which is $r_t \sin(\pi/4)$ as follows:

$$r_c = \frac{1}{r_t \sin(\pi/4)} \int_0^{r_t \sin(\pi/4)} \sqrt{r_t^2 - y^2}\, dy = 0.9087 r_t \tag{2}$$

The reciprocal of the coefficient of r_t is multiplied as a correction factor to obtain the corrected root length (r_t) from the calculated length (r_c). Based on this principle, the length estimator proposed in [12] is:

$$L = 1.2 N_{tot} \times p_d \tag{3}$$

Where, L is the actual length. Similar to this, the length estimator proposed in [23] as:

$$L = 1.1 N_{tot} \times p_d \tag{4}$$

The method proposed in [24] takes into account the number of diagonal and orthogonal connections as:

$$L = \left(\sqrt{2} N_{dc} + N_{oc}\right) \times p_d \tag{5}$$

One serious shortcoming of this equation is that the length is mostly going to be over measured as AB ≤ AD + DB (refer to figure 1a) [13]. Hence the first octant average [25] was introduced as a correction to the bias of equation (5), and the equation becomes:

$$L = 0.948\left(\sqrt{2} N_{dc} + N_{oc}\right) \times p_d \tag{6}$$

The Pythagorean right triangle approach for estimating length could be absolutely error free but will work effectively for straight line segments. For root segments made up of more

than one straight line, which is generally the case, implementation of Pythagorean estimator will be very cumbersome for each line segment. Another length estimator was proposed in [13] with the use of virtual lines as:

$$L = \left[\left\{N_{dc}^2 + \left(N_{dc} + mN_{oc}\right)^2\right\}^{1/2} + (1-m)N_{oc}\right] \times p_d \quad (7)$$

Where, m is a constant ($0 \leq m \leq 1$). The value of m was optimized to an empirical 0.5 based on the simulation results. It can be noted that all the above length estimators rely on the rationale of the skeletonization procedure for binary images. This morphological operation of image analysis has its own problems (refer to the image analysis section for details) and the success of this procedure has serious implications on the actual measured length. In view of this, another length estimator was proposed that makes use of the morphological operation of perimeter determination in lieu of skeletonization [26]. Such an algorithm for root length measurement was earlier proposed [27,28]. Basis of this algorithm is that the length L and width W of a rectangular shape can be related as $2(L+W) = P_{root}$ and $LW = A_{root}$ and any root like structure can be represented as a combination of rectangular shapes in an image. Here P_{root} and A_{root} are perimeter and area of the shapes concerned. Thus the estimator is given by:

$$L = \left[P_{root} + \left(P_{root}^2 - 16A_{root}\right)^{1/2}\right] \times p_d / 4 \quad (8)$$

This has the added advantage of screening off extraneous objects from the root sample image using the length-to-width ratio [7] and thus simplifying the sample processing.

All above mentioned formulae are based on the assumption of random orientation of root samples in the image. This assumption may introduce bias depending on density of root samples [28] or large standard deviation for different orientations of fewer roots [29], but should suffice the purpose as long as density of root samples is fairly large. In case of fewer root samples, Pythagorean estimator can be implemented to minimize errors. With the increase in root density, the overlaps increase proportionally. While determining the orthogonal and diagonal connections, this is calculated as described in [13]. A comparative study of the above length estimators is done to choose the best one for our case.

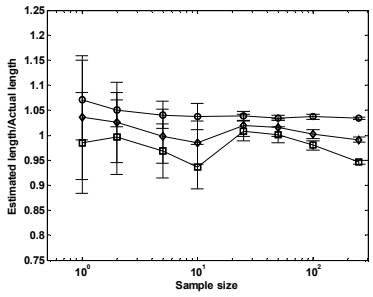


Figure 2. Ratio of measured length to actual length for three estimators as a function of sample size. ◊ – equation (4); ○ – equation (7); □ – equation (8).

A. Comparative performance of length estimators

A simulation program has been developed to generate sets of images of straight lines. The total length of these lines in a given image is known. Sets of images are simulated for different number of straight lines of random orientation and total lengths. These images are subjected to length measurement with estimators represented by equations (4), (7) and (8) respectively. Results are shown in figure 2. It can be seen that there is a dip at 10 lines per image for the estimators of equation (4) and (8). This can be attributed to the overlap effect which is not addressed effectively in these two estimators. The image resolution and size has been increased 25 times for the images beyond 10 lines per image. As the number of lines again increased, another dip can be observed at 250 lines/image. This is because that even at this increased image size, the overlaps again starts to dominate due to the corresponding increment in the number of lines. The estimator of equation (7), on the other hand, remains almost flat even at considerable overlaps as the overlaps have already been taken into account. Hence equation (7) seems to be the most robust estimator albeit a constant offset ~0.04 has been envisaged.

Detail of the second aspect of this paper of leaf area measurement is discussed in the image analysis section.

III. MATERIALS AND METHODS

A. Plant material

Seeds of *A. indica* are sown in Plastic trays filled with sterilized sand: soil mixture (1:2). Seeds germinated in germination tray within 8 days after plantation (DAP) and the seedlings are maintained there for another 3 days. At 11 DAP the seedlings of uniform length (2.5 cm) are selected and transferred from germination tray to root trainers of 150 cc capacity containing the same sand: soil mixture used for raising seedlings. The seedlings are maintained there for 40 days. Such seedlings are inoculated with vermicompost (at 55 DAP; treatment T2) or Arbuscular Mycorrhizal Fungi (AMF; at 40 DAP; treatment T3) or both (treatment T4) for observing the effect of such treatments in this tree species. Treatment T1 is taken as control. At 55 DAP seedlings are transferred from root trainers to polythene bags of 1.5 kg and maintained under ambient condition (12h photoperiod) with normal watering in the net house for the study. Root system from the plants (two plants from each of the four replications of each treatment) is sampled at an interval of 30 days till the end of the experiment (at 150 DAP). Further, all the leaves are collected from the experimental seedlings of all the treatments (similar to root length measurements) for the leaf area measurements.

B. Image acquisition and analysis system

Tender roots of the plant samples are often off white in color and vary in diameter from tens of µm (secondary or tertiary roots) to ~1mm (tap root). This also varies with the age of the plant sample. The roots system along with the soil is collected carefully from the root trainers/ polythene bags and

suspended in water to get rid of the soil. This also ensured minimum damage or loss of delicate root fibers. The root system is further cut into smaller pieces and distributed evenly over a black surface on the base of the image acquisition setup. The root samples are not stained with Coomassie Brilliant Blue as opposed to the conventional technique [23] as the root samples will be required later for further biochemical analysis. This has also proven to be useful as lesser number of plants have to be sacrificed for the experiment. The images are acquired with a digital camera (CANON Powershot A550) equipped with a 7.1 megapixel CCD chip. The color images (RGB) are taken at 1944 (rows) × 2592 (columns) resolution and transferred to the computer in JPEG format. Each image is of the size of 1.2 Mb. Field of view (FOV) of the camera is calibrated with the image of a graph paper, attached on the base of the acquisition setup and thus the spatial resolution of the images at the object plane is ascertained.

For total leaf area measurement, all the leaves from a single plant are collected and scanned with a Hewlett-Packard scanner (hp scanjet 2300c) at 600 dpi. Each image is of 1 Mb. Scanner field dimensions are measured precisely with vernier calipers and thereby the spatial resolution of a pixel is determined. RGB images (7020 × 5100) are preferred to binary images to take into account any change in color of the leaves due to drying. A heat-map image (mean intensity of RGB with 256 intensity levels) is formed from the RGB image. Pixels with intensity less than a threshold value, indicative of the leaves, are counted and multiplied by the area of a single pixel to obtain the total leaf area.

C. *Image* processing

RGB images are represented by a 1944 × 2592 × 3 matrix of 0 to 255 grades in intensity (8 bit). These images are converted into grayscale images of 1944 × 2592 dimension. A gross threshold correction is done to make the background intensity level uniform. This is done by breaking up the image in smaller chunks and thereby setting the intensity level of all the pixels which are less than 10% of the maximum intensity of the chunk to zero. A square shaped morphological structuring element is created with a width of 30 pixels. A morphological opening of the grayscale chunk is performed with the square structuring element. This operation is essentially erosion followed by dilation of the image, using same structuring element for both operations. Thereby, background is subtracted from each chunk and a complete image is reconstructed. Now the grayscale image is converted into a binary image with intensity levels of 0 or 1. The binary image is subjected to a series of morphological operations for smoothing the edges of the roots such that the artifacts of image skeletonization can be avoided later on. This operation may need to be optimized slightly once for a typical set of images. Total root length in the image can be measured from a thinned root image. For that, binary images are subjected to skletonization. Preprocessing procedures for skeleton pruning are optimized in the code.

After skeletonization all root strands are represented by vertical or horizontal (or a combination of both) connections of pixels. For equation (4) length of such a line of pixels can be measured by multiplying the number of pixels by p_d. For implementing equation (7), numbers of orthogonal and diagonal connections are counted by analyzing the eight neighboring pixels for all skeleton pixels. Kernel multiplication is used for this analysis. This is one of the sliding neighbor operations which are used to implement linear filtering using a matrix of integer weights [30].

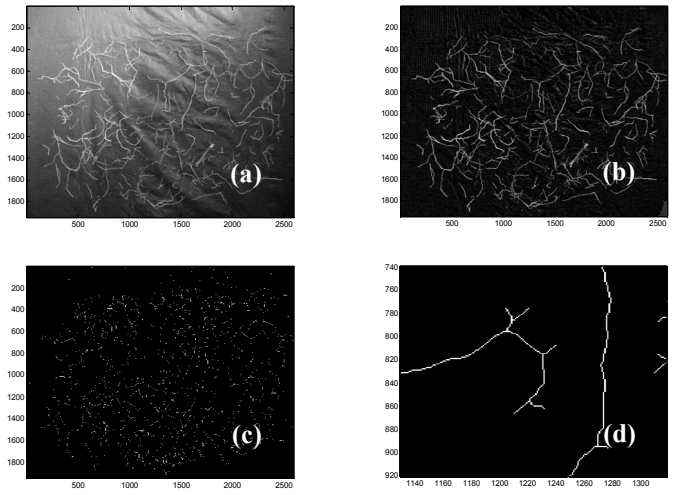


Figure 3. Various stages of the image analysis technique for measuring root length; (a) greyscale image with a gradient in background intensity; (b) same image with uniform intensity background; (c) skeleton image; (d) zoomed view of a portion of the skeleton image to show the one pixel thickness representing the strands of root in the original image.

This operation transforms the value of pixels according to the numbers of orthogonal and diagonal neighbors. Number of pixels that had same value was binned by histogram function and the number of pixels corresponded N_{oc} and N_{dc} were summed up to calculate root length. For implementing equation (8), perimeter of all objects in the binary image is determined and total surface area is also measured. Average root diameter is calculated at this point using this surface area and measured root length. It can be noted, that when images appear to be of low contrast or some extraneous objects are inevitable in the FOV, we have the option of reverting to the procedure of [26], which is more reliable in such cases. The programming platform used here is MATLAB with the image processing toolbox. Student version of MATLAB & SIMULINK R2010a comes with the image processing toolbox and is reasonably cheap. MATLAB being readily available for all platforms like PC, Macintosh or Linux, codes developed for this analysis are also platform independent.

Root image is subjected to the image analysis routine and the actual length is obtained as the output. Different stages of image analysis are shown in figure 3. Figure 4 shows original scanned image (a) and heat-map image (b) of all leaves from a single plant. It has been observed that the area of grey parts of the leaves (as shown in figure 4) is calculated accurately.

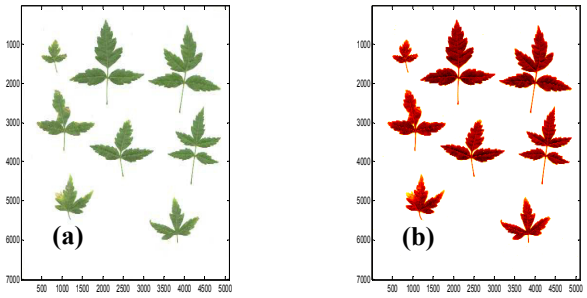


Figure 4. (a) and (b) denote the scanned RGB image of all the leaves from a plant and (b) heat-map image of the same

IV. RESULTS AND DISCUSSION

A. *Accuracy of image analysis*

For further validation of both root length and leaf area measurement methods, the image processing routines are subjected to the measurement of length and area of known samples. For length measurement, a set of images is acquired with known lengths of cotton thread (200 µm diameter) cut into small pieces. Analysis of these images is carried out in the same manner as done for actual root length measurement. Accuracy of these measurements is also compared against the measurements done manually using modified line-intersect method [2]. Results are shown in figure 5. It can be seen that the estimator of equation (7), shown in figure 5c, yields highest correlation coefficient (0.9997) and lowest norm of residuals (112.58) and these are closest to the manual measurements (figure 5a) with a variation of 2-8%. Similar results are obtained when the cotton threads are replaced with copper wire strands (100 µm diameter). Hence, our choice for the estimator suggested in [13] is further validated.

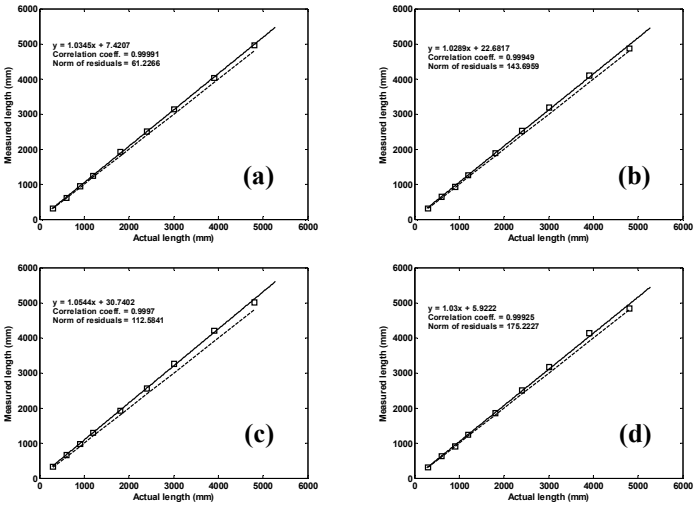


Figure 5. Linear regressions for the actual and measured length of thread samples. (a) to (d) represent the manual line intersect method, equation (4), equation (7) and equation (8) respectively.

For validation of area measurement protocol, coloured papers of known area are taken and cut into pieces of varying shapes and sizes for scanning. Different colours of these papers are chosen to study the implication of any discoloration or drying of leaves on the image analysis. From figure 6 it is evident that the actual area of the coloured pieces of paper could be measured quite precisely with our image analysis technique. High correlation coefficient (0.99997) and low norm of residuals (27.18) are obtained. High resolution (600 dpi) of the scanned images has ensured accurate area measurement with all possible shapes that have been cut out from the paper pieces of known area. This is essential for taking into account the finer details of the leaflets of *A. indica.*

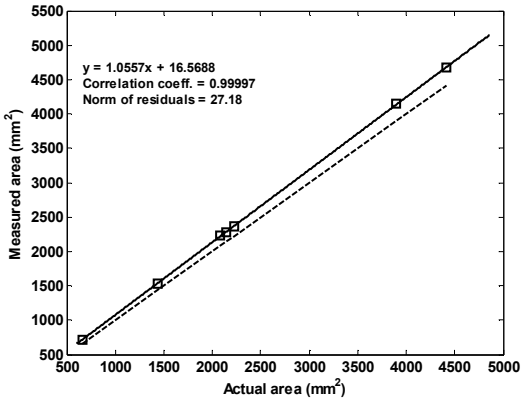


Figure 6. Figure: Linear regression for the actual and measured area for coloured paper samples of different area.

B. *Total root length of Azadirchta indica seedlings*

Root length has been measured for the seedlings at an interval of 30 days from the date of planting (DAP). Eight seedlings per treatment (i.e. two seedlings from each of the four replications per treatment) have been subjected to the root length measurement. Figure 7(a) shows the measured root length as a function of age of the seedling for all four treatments. All treatments have performed significantly better as compared to the control (T1), with T3 being the best.

C. *Total leaf area of Azadirachta indica seedlings*

Figure 7(b) shows the variation of total leaf area as a function of age of the seedling for all the treatments. Total leaf area has increased with time till 120 DAP for T1 and T4. After that a dip in total leaf area has been observed at 150 DAP due to the partial shedding of leaves at the onset of winter. Such an effect was not apparent in case of T2 and T3. All treatments have shown significantly greater leaf area as compared to the control (T1) with T3 being the best. It can be noted that the plants have grown the most in the interval of 60-90 DAP.

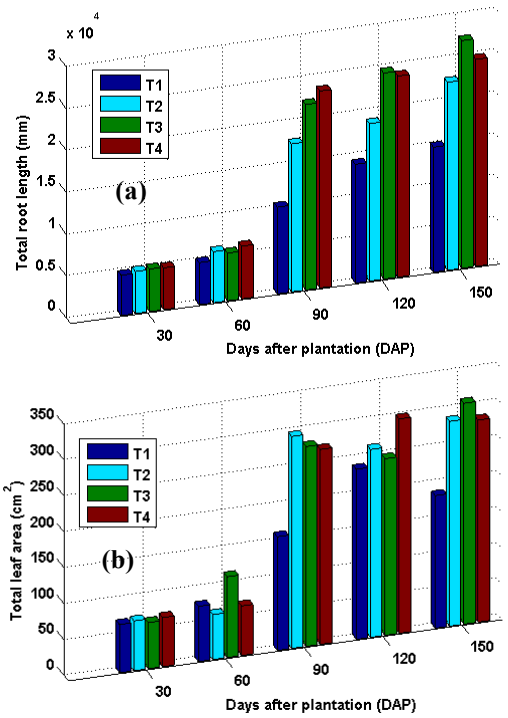


Figure 7. (a) and (b): Total root length and leaf area as a function of DAP.

It can be noted that a fairly high resolution was opted for the root length images. This helps us to sample even the finest root segments distinctly. On the contrary, this high resolution may inflict artifacts in image processing while skeletonization is performed. To alleviate this problem morphological pre-processing operations are performed before the skeletonization and this can be afforded well at this high resolution.

ACKNOWLEDGMENT

The authors would like to acknowledge Dr. H. S. Singh, Mr R. N. Tripathi, Mr R. B. Zala, Mr. M. H. Gadani and Mr. H. R. Parmar for many useful discussions. One of us (K. B.) would like to acknowledge the useful inputs and software support from Mr. Santanu Banerjee.

REFERENCES

[1] E. I. Newman, "A method of estimating the total length of root in a sample," J. Appl. Ecol. Vol. 3, pp. 139-145, 1966.

[2] D. Tennant, "A test of a modified line-intersect method of estimating root length," J. Ecol. Vol. 63, pp. 95–1001, 1975.

[3] W. L. Bland and M. A. Mesarch, "Counting error in the line-intercept method of measuring root length," Plant and Soil vol. 125(1), pp. 155-157, 1990.

[4] D. Richards, F. H. Goubran, W. N. Garwoli and M. W. Daly, "A machine for determining root length," Plant Soil vol. 52, pp. 69–76, 1979.

[5] R. E. Farrell, F. L. Walley, A. P. Lukey and J. J. Germida, "Manual and digital line-intercept methods of measuring root length: a comparison," Agron. J.vol. 85, pp. 1233–1237, 1993.

[6] G. A. Harris and G. S. Campbell, "Automated quantification of roots using simple image analyzer," Agron. J. vol. 81, pp. 935–938, 1989.

[7] S. L. Murphy and A. J. M. Smucker, "Evaluation of video image analysis and line-intercept methods for measuring root system of alfalfa and ryegrass," Agron. J. vol. 87, pp. 865–868, 1995.

[8] W. W. Wilhelm, J. M. Norman and R. L. Newell, "Semiautomated X-Y-plotter-based method for measuring root lengths," Agron. J. vol. 75: 149-152, 1983.

[9] J. J. Krstansky and G. S. Henderson, " Computerized measurement of fine root length using a desktop image scanner" In Agronomy abstracts. pp. 307. ASA, Madison, WI, USA, 1989.

[10] G. Kirchhof, "Measurement of root length and thickness using a hand held computer scanner," Field Crops Res. vol. 29, pp. 79–88, 1992.

[11] W. S. Rasband and D. S. Bright, "A public domain image processing program for the Macintosh," Microbeam Anal. vol. 4, pp. 137–149, 1995.

[12] A. L. Smit, J. F. C. M. Sprangers, P. W. Sablik and J. Groenwold, "Automated measurement of root length with a three-dimensional high-resolution scanner and image analysis," Plant Soil vol. 158, pp. 145–149, 1994.

[13] K. Kimura, S. Kikuchi and S. Yamasaki, "Accurate root length measurement by image analysis," Plant Soil vol. 216, pp. 117–127, 1999.

[14] Xiaojun Qi, Ji Qi and Wu Yajun, "RootLM: a simple color image analysis program for length measurement of primary roots in Arabidopsis," Plant Root vol.1, pp. 10–16, 2007.

[15] E. Rico-García, F. Hernández-Hernández, G. M. Soto-Zarazúa and G. Herrera-Ruiz, "Two new methods for the estimation of leaf area using digital photography," Int. J. Agric. Biol. vol. 11, pp. 397–400, 2009.

[16] E. D. Cittadini and P. L. Peri, "Estimation of leaf area in sweet cherry using a non-destructive method.," RIA vol. 35(1), pp. 143-150, 2006.

[17] O. Kurt, H. Uysal and S. Uzun, "Non-destructive leaf area estimation of flax (Linun usitatissimum)," Pak. J. Bot. vol. 37(4), pp. 837-841, 2005.

[18] Z. Li, C. Ji and J. Liu, "Leaf Area Calculating Based on Digital Image" in IFIP International Federation for Information Processing, Volume 259, Computer and Computing Technologies in Agriculture, vol. 2, Daoliang Li; (Boston: Springer), pp. 1427-1433, 2008.

[19] J.T. Tsialtas and N. Maslaris, "Evaluation of a leaf area prediction model proposed for sunflower," Photosynthetica vol. 46(2), pp. 294-297, 2008.

[20] Matthew E. O'Neal, Douglas A. Landis, Rufus Isaacs, "An inexpensive, accurate method for measuring leaf area and defoliation through digital image analysis," J. Econ. Entomol. vol. 95(6), pp. 1190-1194, 2002.

[21] C.A. Glasbey and G. W. Horgan, "Image Analysis for the Biological Sciences," John Wiley and Sons, Chichester, pp. 164, 1995.

[22] J. Chikushi, S. Yoshida and H. Eguchi, "A new method for measurement of root length by image processing," Biotronics vol. 19, pp. 129-135, 1990.

[23] S. Tanaka, S. Yamauchi and S. Kono, "Easily accessible method for root length measurement using an image analysis system," Jpn. J. Crop Sci. vol. 64(1), pp. 144–147, 1995.

[24] H. Freeman, "Boundary encoding and processing," In Picture Processing and Psychopictorics. Eds B S Lipkin and A Rosenfeld. pp. 241-266. Academic Press, New York, USA, 1970.

[25] Z. Kulpa, "Area and perimeter measurement of blobs in discrete binary pictures," Comput. Vision Graphics Image Process. vol. 6, pp. 434–454, 1977.

[26] T. Vamerali, M. Guarise, A. Ganis, S. Bona and G. Mosca, "Analysis of root images from auger sampling with a fast procedure: a case of application to sugar beet," Plant and Soil vol. 255, pp. 387–397, 2003.

[27] W. L. Pan and R. P. Bolton, "Root Quantification by edge discrimination using a desktop scanner," Agron. J. vol. 83, pp. 1047–1052, 1991.

[28] R. P. Ewing and T. C. Kaspar, "Accurate perimeter and length measurement using an edge chord algorithm," J. Comput. Assist. Microsc. Vol. 7, pp. 91–100, 1995.

[29] F. C. Zoon and P. H. Van Tienderen, "A rapid quantitative measurement of root length and root branching by microcomputer image analysis," Plant Soil vol. 126, pp. 301–308, 1990.

[30] J. C. Russ, The Image Processing Handbook, 2nd ed. CRC Press, Boca Raton, FL, USA, pp. 155, 1994.

Archiving Sensor Data

Applied to Dam Safety Information

José Barateiro[1,2], Gonçalo Antunes[1], Hugo Manguinhas[1], José Borbinha[1]
[1] INESC-ID, Information Systems Group, Lisbon, Portugal
[2] National Laboratory for Civil Engineering, Lisbon, Portugal
{jose.barateiro,goncalo.antunes,hugo.manguinhas,jlb}@ist.utl.pt

***Abstract*—The consequences of structural failures in large civil engineering structures are potentially catastrophic, varying from high economic impacts to unrecoverable environmental damage or loss of life. To prevent that, these structures can be continuously monitored, therefore the management and preservation of the resulting data is crucial to support decisions concerning structural safety. However, preserving data also entails several risks and threats, comprising strong safety requirements. This paper analyzes the scenario of civil engineering safety, presenting the current systems used at the Portuguese National Laboratory for Civil Engineering to manage and preserve sensor data. The main risks that can impede the digital preservation of data are discussed and a solution is proposed where sensor data is objectively described and packaged in order to be reused in the future. This includes controlling the extraction of data from the operational systems, describing the representation of data through a Metadata Registry, and package the context information using a METS aggregator.**

Keywords— Sensor Data; Digital Preservation; Risk Management; Information Management; Workflow.

I. INTRODUCTION

The safety of large civil engineering structures like dams, bridges or nuclear facilities require a comprehensive set of efforts, which must consider the structural safety, the structural monitoring, the operational safety and maintenance, and the emergency planning [1]. The consequences of failure of one of these structures may be catastrophic in many areas, such as: loss of life (minimizing human casualties is the top priority of emergency planning), environmental damage, property damage (e.g., dam flood plain), damage of other infrastructures, energy power loss, socio-economic impact, etc.

The risks associated with these scenarios can be mitigated by a number of structural and non-structural preventive measures, essentially to try to detect in advance any signs of abnormal behavior, allowing the execution of corrective actions in time. The structural measures are mainly related to the physical safety of the structures, while the non-structural measures can comprise a broad set of concerns, such as operation guidelines, emergency action plans, alarm systems, insurance coverage, etc.

In order to improve the structural safety of large civil engineering structures, a substantial technical effort has been made to implement or improve automatic data acquisition systems able to perform real-time monitoring and trigger automatic alarms. This paradigm creates an imminent deluge of data captured by automatic monitoring systems (sensors), along with data generated by large mathematical simulations (theoretical models). Besides the fact that these monitoring systems can save lives and protect goods, they can also prevent costly repairs and help to save money in maintenance. In scenarios like this, it is crucial to provide solutions that support interoperability (i.e., the ability of two or more systems to exchange information and to use the information that has been exchanged [2]), including the concept of temporal interoperability (i.e., long-term preservation).

This paper focuses on the digital preservation dimension of interoperability, which aims to ensure that digital data remain authentic, accessible and understandable over a long period of time. As a first assumption, one can consider that the main reason to preserve data is to preserve its value, as an asset. Consequently, it does not make sense to preserve valueless data. However, to determine and assess the value of data is a difficult and error-prone task. On the other hand, it could be an error to consider that data that cannot be used today will have no value in the future. For instance, today's technology allows the simulation of mathematical models with a much higher resolution and volume of simulated data that was not possible a decade ago.

From this perspective, we assume that the preservation of data concerning the safety of large civil engineering structures is crucial, since: (*i*) observational data is unique and impossible to recreate, (*ii*) complies with legal requirements or contracts established with third-parties, (*iii*) allows the re-use of data for new research, and (*iv*) reduces costs (e.g., the retention of expensively generated data is cheaper to maintain than to re-generate) [3].

The work presented here was developed in the scope of the SHAMAN project[1], which has the aim of developing digital preservation techniques and tools. We analyze the scenario of monitoring dams to assure their structural safety. We show that the digital preservation of sensor data has to deal with the requirements of managing dynamic data, as sensors are continuously capturing data; and heterogeneous and potential large set of representation schemas. Finally, we present an approach, based on the Open Archival Information System (OAIS) Reference Model [4], and a working technical solution

[1] *http://shaman-ip.eu*

implemented specifically to address the challenges of dam sensor data.

The remainder of this paper is organized as follows: Section II describes the scenario of monitoring concrete dams in the scope of the Portuguese National Laboratory for Civil Engineering[2] (LNEC). In section III, the main risks that can hamper the preservation of this information are discussed. Section IV describes the proposed solution to digitally preserve dam sensor data. Finally, Section V resumes the main conclusions and future work.

II. DAM SAFETY

The interpretation of the correlation of several parameters measured, in different physical locations of a structure, can be used to validate the current state of that structure and predict its future behavior under specific and controlled conditions [1]. This is a key factor to detect potential anomalies and to be able to make decisions on time, reducing the risk of failures with catastrophic consequences. In the case of concrete dams, for example, their behavior is continuously monitored by instruments (e.g., plumb lines, piezometers) installed in strategic points of the dam [5] [6], which can typically range from hundreds to few thousands of instruments or sensors.

The related raw data, usually known as "readings", is collected manually by human operators or collected automatically by sensors. These readings are transformed, by specific algorithms, into engineering quantities (physical actions that can be used to assess the behavior of the structure as, for example, a tension or a relative displacement). Actually, the term "reading" does not clearly correspond to raw data, since a reading is already a transformation from the raw data. For instance, an electrical instrument like an extensometer might provide raw data as a voltage (mV), which is then converted by a reading instrument (or by the sensor) into a resistance and a resistance relation, which are finally converted into an extension (engineering quantity). This monitoring information includes, essentially, instrument properties, readings and engineering quantities.

The Portuguese regulations [7] state that the National Laboratory for Civil Engineering is responsible for keeping an electronic archive of data concerning the dam safety. Thus, the preservation of this data is a legal obligation. Moreover, that obligation defines the duties of the different parties involved in dam safety, namely the dam owners, the dam safety authority and the dam engineers and builders. As a consequence, several entities are compelled to share data, and thus must face interoperability and preservation issues when dealing with heterogeneous sources of information [8].

Currently, LNEC uses a modular information system (*GestBarragens*) that provides components to manage dam observations, visual inspections, physical models and mathematical models. It also supports the management of technical documents and provides a set of exploitation tools, in the form of tabular and chart reports, graphical visualization of geo-referenced information, among others. However, the *GestBarragens* system was not designed for preservation purposes. Indeed, it supports the operational procedures to manage information concerning the dam safety, but does not assure the preservation of this information. It is a web-based system developed on the top of the .NET framework, where the underlying data is stored and managed in an Oracle 10g database. It uses a SOAP interface to provide and expose exploitation services as well as multiple ingest services.

TABLE I. summarizes an example of the data concerning the dam safety of a concrete dam. Currently, LNEC supports 32 different types of instruments with manual data acquisition and 25 different types of automatic monitoring instruments (implemented with sensors). Both the number and type of instruments installed in a specific structure depend on the stage of the structure's life and on a few hundred to thousand specific parameters that affect its behavior. Currently LNEC monitors about 80 concrete dams, generating an average of 264,000 records per day that have to be processed and preserved.

III. DIGITAL PRESERVATION RISKS

Although it is impossible to define all the requirements applicable for all digital preservation needs, a survey was made following a set of requirements based on the scenario presented in Section II.

First of all, digital preservation requires that a copy (or representation) of any preserved digital data survives over the actual system's lifetime, which is usually unknown, but may be as long as decades or even centuries (LNEC monitors concrete dams of more than 80 years old). This can be defined as a **reliability** requirement. Therefore, a digital preservation system must be designed to preserve data for an indefinite period of time without suffering any data losses.

Also, a future consumer should be able to decide if the accessed information is sufficiently trustworthy. Usually, this requires the assurance of the **authenticity** of digital data (which is already a common requirement for tangible objects), along with an accurate identification of their **provenance** (typically information about its creation, responsible entity, lineage, etc.). Moreover, it is crucial to assure the **integrity** of digital data, guaranteeing that their information content was not modified. Authenticity, provenance and integrity are thus crucial requirements for qualified specialists to trust and correctly approximate and estimate the behavior of large civil engineering structures.

The provenance requirement is fundamental. Furthermore, the complex scientific computations that occur in the production workflow (e.g. calculation of engineering quantities from raw data, outlier's detection) also make it complex to manage. The production workflow can be seen as an example of a scientific workflow [9] where data transformations and analysis steps, as well as the mechanisms to carry them out, are captured and represented as a workflow [10]. An access and re-use scenario, which is common in the simulation of mathematical models should also be considered. In fact, a mathematical model consumes observational data (preserved in the archive) and produces new digital data that should be preserved.

Third, digital preservation requires that future consumers are able to obtain the preserved information as its creators

[2] *http://www.lnec.pt*

TABLE I. TYPICAL DATA REGISTERED FOR A REPRESENTATIVE CONCRETE DAM

Data Stage	*Description*	*# per day*	*Format*	*Notes*
Raw	Depend on the instrument type (e.g. voltage)	Currently discarded	Proprietary to the sensor	This information is currently discarded by sensors and not registered during manual acquisition
Processed readings	Transformed from raw data	Aprox. 3300 rows	.xls, .mdb, PDT, ascii	Sensors register data in .xls or .mdb and access a web service to send this information to LNEC. Manual acquisition can be registered into a PDT and automatically sent to LNEC or inserted via web interface or text file
Calculated engineering quantities	Calculated from readings	Aprox. 3250 rows	Oracle database	Algorithms to filter, clean and calculate engineering quantities are implemented as Oracle stored procedures (PL/SQL)
Analyzed	Tables, graphs, gis, mathematical simulations	Varies	.html, .xls, .pdf, .dxf (CAD), .xml	Uses several tools, including reporting tools and a geographic information system

intended, thus it must **deal with obsolescence** threats [11]. This requirement encloses several challenges, since digital data to be explored, require a technological context defined by specific software and, in some cases, even by specific hardware [12]. Moreover, in this special scenario, it is also crucial to preserve the processes involved in the creation of the preserved data. For instance, the scientific workflow for data acquisition must be preserved and linked with the generated data.

Finally, dynamic collections and environments for digital preservation require technical **scalability** to face technology evolution allowing, for instance, the addition of new components through incremental updates [13]. This also implies a requirement for supporting **heterogeneity** (which is reinforced by the requirements for scalability).

In previous work, a taxonomy for digital preservation risks (see TABLE II.) was proposed, which considers that a risk is the impact that occurs when an event (threat) is able to exploit a system vulnerability, affecting the achievement of the digital preservation requirements described above.

TABLE II. THREATS TO DIGITAL PRESERVATION [14]

Vulnerabilities	Data	Media faults Media obsolescence
	Process	Software faults Software obsolescence
	Infrastructure	Hardware faults Hardware obsolescence Communication faults Network service failures
Threats	Disasters	Natural disasters Human operational errors
	Attacks	Internal attack External attacks
	Management	Economic failures Organization failures
	Business	Legal requirements Stakeholders' requirements

Like common information system's architectures, this paper considers a preservation environment as the aggregation of different components, namely: (*i*) the information entities, including preserved digital data and metadata, (*ii*) processes controlling the information entities (can be supported by computational services), and (*iii*) the technological infrastructure that supports the preservation environment.

Each of these components may present several vulnerabilities, which we classify as: (*i*) data vulnerabilities, affecting the information entities, (*ii*) process vulnerabilities, affecting the execution of processes (manual or supported by computational services) that control information entities, and (*iii*) infrastructure vulnerabilities, enclosing the technical problems in the infrastructure's components.

A classification of threats to digital preservation is also proposed which distinguishes threats into four categories: disasters, attacks, management and business. Disasters and attacks correspond, respectively, to non-deliberate and deliberate actions affecting the system or its components. Management failures are the consequences of wrong decisions that produce threats to the preservation environment. Finally, business threats depend on a specific business context and occur when new or updated legislation, as well as new or updated requirements defined by related stakeholders concerned with the business, can produce an impact on the achievement of digital preservation requirements.

Some risks can remain unnoticed for a long period of time. For instance, a damaged hard disk sector can remain undetected until a data integrity validation or hard disk check is performed. Furthermore, one cannot assume threat independence, since a specific threat can generate other threats.

Considering the risks to digital preservation, this paper claims that a "digital preservation system" is itself an infrastructure in risk, comprising strong safety requirements, as happens in civil engineering structures. Moreover, since the safety of large civil engineering structures is directly dependent to the monitoring systems and the preservation of the associated data, the consequences of a failure in the preservation system can also produce catastrophic effects (e.g. loss of life, environmental damage, etc.).

IV. SOLUTION OVERVIEW

When addressing the problem of digital preservation for memory institutions (e.g., libraries, archives, museums) where the digital data to be preserved are typical static documents (e.g., images, text documents), it is a common accepted solution to apply the OAIS reference model, since the information package is composed by the digital data and a set of metadata associated with them. The ultimate objective of solutions based on the OAIS reference model is to mitigate the risks identified in Section III.

Research undertaken in the SHAMAN project determined that a bigger understanding of the context surrounding the production, preservation, and reuse of information (OAIS view) was needed in order to understand its implications on preservation. Thus, a model of the lifecycle of information was created and can be seen in Figure 1.

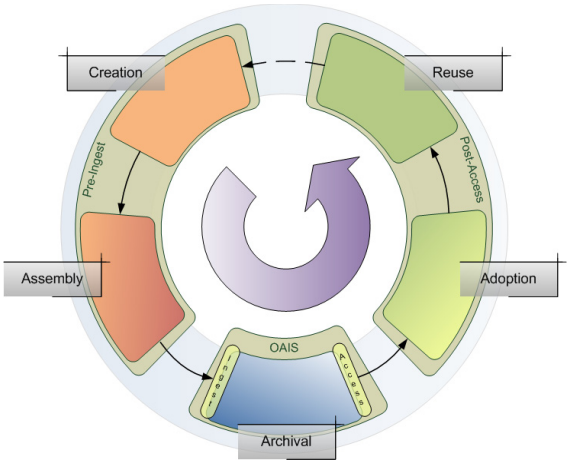


Figure 1. Information lifecycle [15]

During the *creation* phase new information comes into existence. Normally, information is not created for the purpose of archiving, thus archivable information can be the result of complex processes that involve a multitude of stakeholders. The *assembly* phase deals with the appraisal of information relevant for archival and all processing and enrichment for compiling the complete information to support future reuse. Normally, this compilation is called an *archival package*. The *archival* phase corresponds to the OAIS reference model and addresses the life-time of the digital data inside the archive, including the ingestion of and access to information. The *adoption* phase encompasses all processes by which information provided by the Archive is screened, examined, adapted, and integrated for the proper reuse. Finally, the *reuse* phase deals with the exploitation of information in the interests of the consumer.

The preservation of dam safety sensor data raises several challenges because of the data and process characteristics. First, data is not static (a data set is continuously increasing). Second, since new sensors (with different characteristics and results) have to be accommodated in the future, new data representations must be handled. Third, the representation of a dataset can evolve in the future (new devices can use different representations to store the same data), limiting the ability to understand the same type of data, as well as relating the same type of data when it was captured by devices using different data representations. Finally, the nature if complex and interlinked objects composed by datasets and their representation (an isolated dataset is useless to interpret the structural behaviour).

In order to control the complexity of data representations, some communities developed their own metadata initiatives as, for instance, the *Ecological Metadata Language (EML*[3]*),* or the *Federal Geographical Data Committee (FGDC*[4]*).* Yet, there will never be a unified metadata schema for all possible data. Thus, in a scenario that is not covered by current metadata initiatives, or when the information can be represented in heterogeneous schemas that can continuously change (like the sensors used in the civil engineering domain), the use of standard languages to describe data representations [16] is an expected solution.

The SHAMAN project developed an archival infrastructure that follows the OAIS reference model and uses the iRODS[5] data grid as storage substrate. The work presented here relies on this infrastructure to address the digital preservation risks related to media faults, process and infrastructure vulnerabilities, as well as the issues related to the volume of data and its imminent deluge. However, in the dam safety context, an information package is an interlinked object that must aggregate the sensor data, the information on the sensors that produced the data, as well as the description of the schemas used to encode them (it can include syntactic and semantic representation). These activities are part of the *creation* and *assembly* phases, which will influence the future adoption and reuse of this information.

The proposed solution elaborates on the creation and description of information packages to control the media obsolescence vulnerabilities that occur when the representation format becomes obsolete and unable to be rendered, even if the "bit stream" survives over time. Since the information package is composed by sensor data (from distinct types of sensors), along with their contextual and representation information, a network of object have to be aggregated to create a meaningful object in the context of civil engineering. The Metadata Encoding and Transmission Standard (METS[6]) is a widespread metadata representation to encode structural metadata in XML. The use of METS provides and extensible way to represent the aggregations required by the illustrated scenario.

On the other hand, to address the management of schema representations (including the definition of sensors, raw data, processed readings, etc.) and their dynamic nature (new or updated schemas to represent the same information), it is critical to manage metadata that describes the information representation. This is not a new requirement in the community, where, for instance, previous work developed the *Metacat* framework [17], which is able to store, retrieve and transform XML documents managed stored in a relational database. In this paper, we use the concept of Metadata

[3] *http://knb.ecoinformatics.org/software/eml*
[4] *http://www.fgdc.gov*
[5] *https://www.irods.org*
[6] *http://www.loc.gov/standards/mets*

Registry (MDR), which was conceived to represent a system that allows the management of multiple schemas (not limited to XML) and the export of information about the schema. It also supports the creation and management of mappings between different schemas. This concept is formalized by the ISO 11179[7] series of standards. Accordingly, the MDR can be used to address the challenges of representing the encoding of sensor data (including the definition of sensors and the data stages listed in TABLE I. also supporting the future migration of information packages.

For demonstrating the preservation of data in this particular scenario, we developed a *Service Oriented Architecture (SOA)* solution, as shown in Figure 2. Our proposal comprises services for: acquiring data stored in the *GestBarragens* information system, acquiring a description of the schema representation, packaging the data together, and ingesting the data package into the archival system. Such a solution is controlled by a service orchestrator (Service Orchestration component) parameterized in *Business Process Execution Language (BPEL)* and executed by a *GlassFish Open ESB BPEL* engine. This way, the BPEL file representing the creation/assembly process of the information package is itself part of the package, which is critical for provenance purposes.

The following components implement *Java Web Services* that are orchestrated by the Service Orchestration component:

- **Data Extractor**: Extracts data from the *GestBarragens* system, according to the parameters defined by the *Assembly Orchestration.* To support the dynamic nature of sensor data, it has the option to define the time window for data extraction, full extraction, incremental extraction and the list of dams to extract. The recursive use of full data sets uses more space, while incremental data sets require the recomposition of data sets on access.
- **Metadata Registry** (MDR): Supports the registration and management of multiple data schemas, addressing "the semantics of data", "the representation of data", and "the registration of the descriptions of that data". It also supports the creation and management of the mappings between data schemas, as well as the export of both schema and mapping information.
- **Data Aggregator**: the METS schema is used to "wrap" all the information, acting as structural metadata. The information that is aggregated by the METS includes: (*i*) data about the characteristics of the sensor which produced the readings (e.g., calibration constants, validation intervals, etc.), (*ii*) schema information of the data containing the characteristics of sensors, (*iii*) observational data, (*iv*) schema information of the observational data, (*v*) BPEL file representing the assembly process, and (*vi*) generated HTML files to facilitate human navigation under the METS components.

The *Assembly Orchestration* component starts by (1) acquiring observational data and sensor information from the *GestBarragens* through the Data Extractor component, specifying both the type of export (time window, full, or

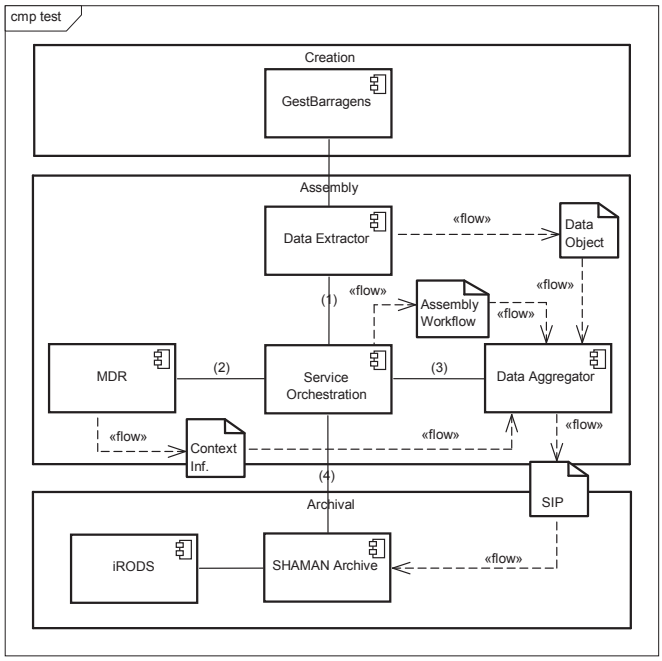


Figure 2. Overview of the proposed solution for digital preservation integrated with the GestBarragens information sytem.

incremental from last export) and the dam or list of dams, and (2) acquiring the related schema information by requesting it from the MDR, depending on the type of data exported in the previous step. It continues, by (3) requesting the generation of a METS file to package the dam information and (4) submitting the package into the archive (using the *ingest Web Service* of the SHAMAN archive). When the submitted package enters the SHAMAN archive, it is then managed as common information packages, as those constructed for typical data objects like images or text. In the case of the SHAMAN archive, an information package is encoded in plain zip and includes an OAI-ORE manifest[8] to aggregate resources contained in the information packages (e.g., information content, preservation metadata).

When the dam safety data is accessed from the archive for future use (*adoption*), the information package is self-contained, in the sense that it includes, not only the preserved data, but also all the information required to render this data (structural information provided by the schema representation extracted from the MDR), in addition to the context information required to understand the data itself (context information like the type and characteristics of sensors, location, data units, etc.).

Finally, the integration of a MDR and the decoupling between data, its schema representation and the context information in the information packages, support the use of migration techniques inside the archive. In fact, migration is one of the most effective techniques used in digital preservation to avoid the obsolescence of data representations/formats. For observational data, mappings between schemas supported by the MDR provide an effective

[7] *http://metadata-standards.org/11179*

[8] *http://www.openarchives.org/ore*

tool to migrate from an obsolete schema representation to an updated representation. Note that migrations can be often lossy, making it critical to plan when, how and what to migrate [18].

The use of a SOA architecture, and the respective service independence, allows the adoption of this solution to several scenarios. From the proposed services, only the *Data Extractor* is scenario dependent. It is also independent from the archival solution. It only requires an archival service that can be accessed through a *Web Service* that can be configured and called by any BPEL engine.

V. Conclusions and Future Work

This paper is motivated by the real case study of managing data concerning the safety of large civil engineering structures. It describes the technological solutions that are being used in LNEC and shows that these solutions were designed for operational purposes and do not address emergent digital preservation requirements. For instance, in this type of scenario, future research requires details about provenance and production workflows (e.g. conversion from raw data to engineering quantities) that are not currently handled. From the analysis of scenarios handled at LNEC, this paper motivates the need for digital preservation and surveys its main requirements and threats. This analysis shows that a preservation system is itself an infrastructure in risk, requiring continuous actions to be safe for a long period of time. Since the safety of civil engineering structures is directly connected to the underlying monitoring data, a failure in the preservation system can potentially produce catastrophic consequences.

In the scope of the SHAMAN project, a digital archive supported by an iRODS data grid infrastructure was developed. This archive adopts the widely accepted OAIS reference model, where digital data is packaged with metadata to be reused in the future. This paper motivates the need to extend this model in order to support complex and dynamic digital data, as those generated by sensors that continuously monitor the behavior of the dam structures. The proposed solution comprises components to manage the complexity of heterogeneous schemas, control the dynamic behavior of datasets, and aggregate context information in a meaningful way. The assembly of this information is also a complex process, which motivates the control of this process through a service orchestrator.

The problem of preserving sensor data differs from the preservation of traditional documents, mainly because sensor data is dynamic and might have heterogeneous data representations. Moreover, the context information is much more complex, since sensor data depends on the sensor properties, calibration constants, etc.

The proposed solution, adopted in the scope of the SHAMAN project, is able to preserve sensor data. However, as future work, the dependencies of sensor data both in the infrastructure, but also on the acquisition process, require the problem to be addressed from the perspective of the overall business process, instead of a data-centric approach. This is the vision proposed by the TIMBUS [9] project, where digital preservation is seen as a business continuity issue, where business processes that ran in the past should be able to be reproduced in the future. In the case of sensor data, the recreation of the overall production environment (to simulate the sensor data acquisition) can be used to study the behavior of structures under a controlled (simulated real) environment.

[9] *http://timbusproject.net*

Acknowledgment

This work was supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds and by the projects SHAMAN and TIMBUS, partially funded by the EU under the FP7 contracts 216736 and 269940.

References

[1] M. Wieland and R. Mueller. Dam safety, emergency action plans and water alarm systems. International Water Power & Dam Construction January 2009

[2] IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries, 1990.

[3] P. Lord and A. Macdonald. E-Science curation report – Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision, 2003.

[4] Consultative Committee on Space Data Systems, ISO 14721:2003 - Reference model for an open archival information system (OAIS), 2003.

[5] ASCE – American Society of Civil Engineers. Guidelines for instrumentation and measurements for monitoring dam performance. 2000. ISBN 0-7844-0531-X.

[6] ICOLD - International Commission on Large Dams. Guidelines for automated dam monitoring systems. 1999.

[7] RSB. Dam safety regulation, DL n.344/2007, October 15th. DR, Lisbon, 2007 (in Portuguese).

[8] M. Franklin, A. Halevy and D. Maier. From databases to dataspaces: a new abstraction for information management. SIGMOD Record, 34(4):27-33, 2005.

[9] B. Ludascher, I. Altintas, C. Berkley, D.Higgins, E. Jaeger, M. Jones, E. Lee, J. Tao, Y. Zhao. Scientific Workflow Management and the KEPLER System. March, 2005.

[10] I. Altintas. Lifecycle of Scientific Workflows and Their Provenance: A Usage Perspective. IEEE Congress on Services. 2008.

[11] J. Barateiro, G. Antunes, M. Cabral, J. Borbinha, and R. Rodrigues. Using a GRID for digital preservation. International Conference on Asian Pacific Digital Libraries, Bali, Indonesia, December 2008.

[12] J. Borbinha. Authority control in the world of metadata. Cataloging Classification Quarterly, 38 Issue: 3/4, 2004.

[13] A. Chervenak, I. Foster, C. Kesselman, C. Salisbury, and Tuecke. The data grid: Towards an architecture for the distributed management and analysis of large scientific datasets. Journal of Network and Computer Applications, 23:187-200, 2000.

[14] J. Barateiro, G. Antunes, F. Freitas and J. Borbinha. Designing Digital Preservation Solutions: A Risk Management-Based Approach. International Journal of Digital Curation 5(1):5-17, 2010.

[15] H. Brooks, A. Kranstedt, G. Jaschke and M. Hemmje. Modeling context for digital preservation. Smart Information Knowledge Management, 197-226, 2010.

[16] M. Westhead, T. Wen and R. Carroll. Describing data on the GRID. GRID Computing, 134-140, 2003.

[17] M. Jones, C. Berkley, J. Bojilova and M. Schildhauer. Managing Scientific Metadata. IEEE Internet Computing 5: 59-68, 2001.

[18] C, Becker, H. Kulovits, M. Guttenbrunner, S. Strodl, A. Rauber, and H. Hofman. Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans. International Journal on Digital Libraries (IJDL), 10(4):133-157, December 2009.

Sensor lifecycle management using scientific workflows

Derik Barseghian[1], Daniel Crawl[2], Matthew B. Jones[1], Ilkay Altintas[2], Jing Tao[1], and Sean Riddle[3]
[1] National Center for Ecological Analysis and Synthesis, University of California Santa Barbara
[2] San Diego Supercomputer Center, University of California San Diego
[3] University of California Davis
{barseghian, jones, tao}@nceas.ucsb.edu, {crawl, altintas}@sdsc.edu, swriddle@ucdavis.edu

***Abstract*—Sensor networks are increasingly being deployed to create field-based environmental observatories. As the size and complexity of these networks increase, many challenges arise including monitoring and controlling sensor devices, archiving large volumes of continuously generated data, and the management of heterogeneous hardware devices. This paper presents the Kepler Sensor Platform, an open-source, vender-neutral extension to a scientific workflow system for full-lifecycle management of sensor networks. This extension addresses many of the challenges that arise from sensor site management by providing a suite of tools for monitoring and controlling deployed sensors, as well as for sensor data analysis, modeling, visualization, documentation, archival, and retrieval. An integrated scheduler interface has been developed allowing users to schedule workflows for periodic execution on remote servers. We discuss and evaluate the scalability of periodically executed sensor archiving workflows that automatically download, document, and archive data from a sensor site. We conclude by discussing and comparing the Kepler Sensor Platform to related software.**

Keywords—sensor network; scientific workflow; data discovery; data preservation; data analysis; quality assurance

I. INTRODUCTION

Automated sensing is increasingly used within field-based environmental sciences that traditionally used much more labor-intensive processes to collect data. In addition to the well-known, large-scale observatory programs (e.g., the National Ecological Observatory Network), individual graduate students, technicians, postdoctoral fellows, and faculty are increasingly specifying, deploying, maintaining, and managing sensor networks consisting of tens to thousands of sensors. These individual researchers face all of the management burdens that these complex, technological systems engender, but have few open software choices available to use in facing these burdens.

Some of the challenges that arise include: 1) the need to manage large volumes of data on a continuous basis; 2) quality assurance analysis for these data streams; 3) archival of both the raw data streams and quality-corrected derived data products; 4) visualization of the data; 5) monitoring of large collections of sensors spanning multiple vendors, each with their own vendor-specific control software; and, 6) control and configuration of these sensors that span vendors. For typical scientific users that have minimal background in technology and programming, these challenges impede their ability to deploy and utilize small to large-scale sensor networks, and therefore limit the effectiveness of these systems for environmental science.

Vendor-neutral tools that assist the user-scientist throughout the lifecycle of sensor data are needed for designing, configuring, deploying, managing, and consuming data from these networks, as well as for monitoring and controlling the deployed sensor networks. Such management tools need to be able to manage sensor networks in many different deployment topologies, and manage and visualize both small and large deployments across sensor manufacturers.

Scientific workflow systems [1], [2] provide tools for authoring, executing, documenting, and archiving analysis and modeling processes. Tools such as Kepler [3] can be used to model many data processing tasks in an intuitive way by visually depicting the graph of steps in any scientific analysis. In previous work, Barseghian et al. [4] showed that scientific workflow systems like Kepler could be used to conveniently access sensor data from common sensor network middleware platforms such as DataTurbine [5]. However, this approach only partially solves the challenges facing scientists trying to manage sensor networks; complete solutions would address management of the full lifecycle of a sensor network, spanning both the systems engineering aspects of the lifecycle (e.g., network design, deployment, configuration, inventory, monitoring, and visualization) and the scientific use aspects of the lifecycle (e.g., data stream consumption, quality assurance, analysis, modeling, documentation, archiving, and visualization).

The major contributions of this work are to describe and evaluate the Kepler Sensor Platform, an extension of a scientific workflow system for full-lifecycle management of sensor networks. The work demonstrates the utility of the workflow system for graphical sensor site management, visualization, and analysis, as well as end-to-end management of sensor infrastructure, from sensors to data archives. The system provides a vendor-neutral client-side sensor management application to handle the sensor engineering lifecycle, and a suite of analysis, modeling, and visualization

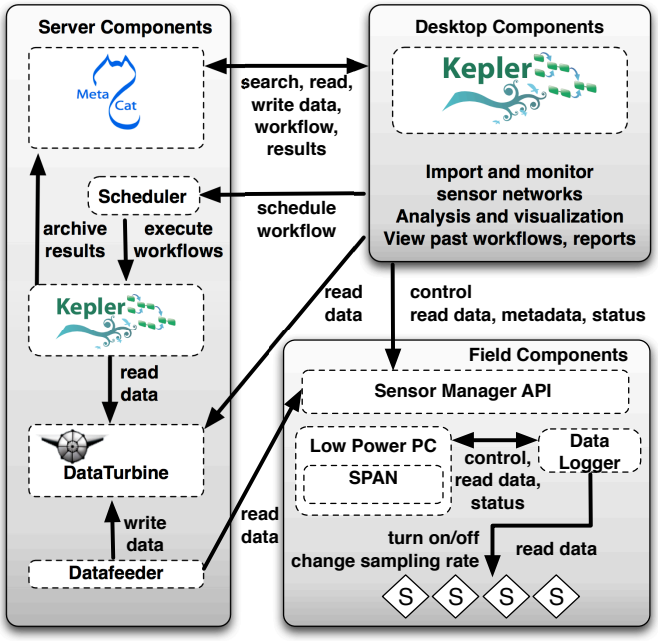


Figure 1. Architecture of the Kepler Sensor Platform. Users interact with sensors using the desktop Kepler application where they can import, layout, and visualize sensor networks, monitor the status of sensors, and get real-time data visualizations. From Kepler, they can schedule archiving and quality assurance workflows that periodically process data from the sensor network, provide metadata, and archive segmented snapshots of the data in the Metacat data archive.

tools to handle the scientific lifecycle. Together, these subsystems integrate sensor management with scientific analysis and modeling systems via the workflow paradigm in a visually intuitive and extensible manner.

In section II, we describe the systems across the lifecycle of sensor data, including the systems engineering aspects of the engineering lifecycle and the use of sensor data in the scientific lifecycle. We evaluate the scalability of the system for typical sensor loads in section III, discuss related work in section IV, and discuss conclusions and future work in section V.

II. Lifecycle of Sensor Data

To effectively manage sensor networks, we designed the Kepler Sensor Platform to provide features targeting both the systems engineering portion of the lifecycle, focused on design, deployment, and monitoring of sensor networks, and the scientific usage portion of the lifecycle, focused on access to sensor data for analysis, modeling, and visualization. Figure 1 shows the main components of the Kepler Sensor Platform, including Field Deployed Components (directly interfacing with sensors), Server Deployed Components (to provide archival and automated processing systems), and Desktop Components (that provide Kepler as a client user interface to the other system components). Communication with the Field Deployed Components is handled through a Sensor Manager interface; this abstraction supports different types of hardware from various vendors. The Sensor Manager communicates with a SPAN (Sensor Processing and Acquisition Network) server that provides drivers and a control interface for each of the sensors in the network [6]. Each of these components is used in both the engineering lifecycle and the scientific lifecycle of sensor data. For example, the server deployment includes components to transfer data from the field Sensor Manager to a DataTurbine server, and a Workflow Scheduler to manage and execute workflows on a Kepler execution engine, which is used to execute a workflow that segments the sensor data and metadata from DataTurbine and archives these to a Metacat data repository [7].

A. *Engineering Lifecycle (Sensor Site management)*

To manage sensor networks, scientists need to be able to design, inspect, monitor, and control suites of sensors deployed in the field. The Kepler Sensor Platform supports these functions through a client-side graphical interface to visualize a sensor deployment site as a workflow using the Kepler GUI (Figure 2). Hardware components such as sensors, and dataloggers can be dragged-and-dropped onto the canvas and connected to one another to represent the actual hardware configuration. Users can provide metadata such as make, model, location, and firmware for each of the hardware components. The canvas may also be annotated with lines, shapes, and text to further document the deployment site. This can be used to convey contextual information about a site, for example to depict spatial layout of sensors, experimental treatments, relevant geographic features, and obstacles like locked gates. Further, an engineering workflow can be exported to KML and viewed in Google Earth to display a satellite view of site components.

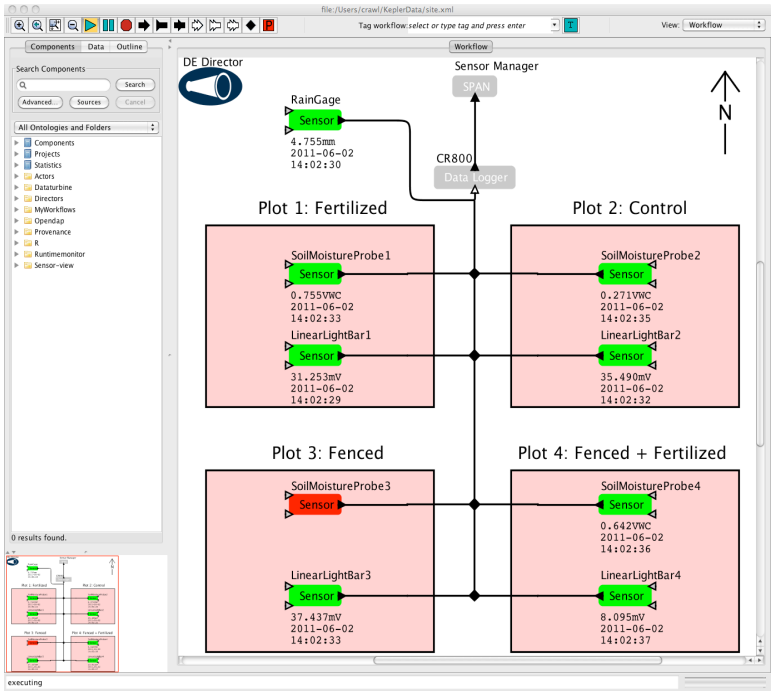


Figure 2. A sensor deployment site in the Kepler Engineering View. Nine sensors connected to a datalogger depict their relationship to real-world experimental plots. Right-clicking on sensors allows their metadata to be viewed and edited. Current sensor values display below the icons, and sensors that are inoperable are shown in red.

Importing sensor sites. To efficiently create engineering workflows, the sensor site description may be imported from a Sensor Manager. In this case, the user need only provide the URL of a Sensor Manager. The Kepler Engineering View queries the hardware descriptions from the Sensor Manager and automatically populates the canvas with components representing the site. The user may then add annotations or edit existing metadata parameters.

Sensor Monitoring. In addition to describing a sensor network site, the Kepler Engineering View provides an interface to monitor deployed hardware components. When an engineering workflow is executed, the Kepler Client queries the sensors' status from the Sensor Manager. As shown in Figure 2, the icons for each hardware component change color based on their status: green for on, red for off, blue for changing parameters, etc. Additionally, for each active sensor, live data values and their associated timestamps are displayed below the icon. Live data may also easily be plotted to show changes over time and to compare data from different sensors.

Sensor control. The Kepler Engineering View allows users to turn a sensor on or off, or change its sampling rate. As described previously, a user can also edit a sensor's metadata parameters. Additionally, a sensor may be controlled using its sensor actor in a scientific workflow (Figure 3). In Kepler, actors read inputs, perform a task, and write outputs. Actors can be connected so that the output of one is read in as input to another. A sensor actor may accept two inputs: sampling rate and a boolean value indicating if it should be active. When a sensor actor executes, it reads these inputs and, if the values have changed, communicates them to the Sensor Manager, which in turn enacts the changes at the site. If active, a sensor actor also outputs the last data value sampled through its output port. This is a powerful feature that allows users to design workflows to monitor sensor values and control sensor sites. A workflow can make changes to sensors based on previous data values from the same or other sensors, creating feedback loops that can be used for adaptive, event-driven sampling. Such adaptive workflows may also contain more involved analyses, e.g., comparison of live data against archived datasets.

Sensor data archiving. Archiving data from sensor networks can be tedious and data loss is difficult to avoid. One challenge is that data collection is continuous, which stresses existing systems that are more transaction oriented. The Kepler Sensor Platform system solves this problem by providing a server-side temporary storage buffer (DataTurbine, an open-source streaming middleware application that provides network ring-buffers for data storage [5]) to reliably accumulate sensor observations and multiplex data from all sensors at a site. The Sensor Manager stores sensor data to DataTurbine as a reliable, short-term cache of the data.

DataTurbine's ring-buffer is necessarily finite in size, so the data must also be archived for permanent long-term storage. For long-term storage, the Kepler Sensor Platform segments each of the data streams into a consistent size, generally based on temporal or spatial windows, generates detailed metadata describing that segment of data, and archives the segment in a Metacat server [7]. Metacat provides a federated storage

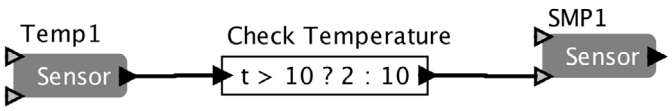


Figure 3. A scientific workflow that contains two sensor actors. In this example workflow, the sampling rate of one sensor is changed based on the output of another. The sensor Temp1 measures the temperature, and SMP1 measures soil moisture. When the temperature measured by Temp1 rises above 10 C, the sampling rate of SMP1 is decreased to 2 Hz; when the temperature goes below 10 C, the sampling rate is increased to 10 Hz.

solution for the Knowledge Network for Biocomplexity (KNB) data federation, and can be used as Member Nodes in the DataONE network [8], thereby making it easy for users to connect the Kepler Sensor Platform to national and international data federation initiatives.

This archival process is accomplished by executing a Kepler workflow that can inspect the sensor site metadata to determine appropriate archiving intervals, connection parameters, and other necessary metadata. The archival workflow compares the currently available data against previously archived data segments, and when appropriate intervals have been reached, automatically downloads, documents, and archives a new segment of the data. A metadata document is created for each data segment. This document is an instance of Ecological Metadata Language (EML) [9], and provides information to describe the data segment, such as sensor and site name, geographic location, temporal period of data collected, measurement units, and a link to access the data. The metadata document also contains a SensorML [10] description of the sensor metadata, such as the device type and manufacturer, and other relevant sensor metadata. The data segments and associated metadata documentation are stored in data packages on the Metacat server. These packages can later be searched for and retrieved with Kepler and data tools like Morpho [11], and via the web.

Workflow scheduling. A workflow scheduler was created to automate the process of archiving sensor data periodically. A user can specify the start time, end time, and execution interval at which the archival workflow should be run. This schedule is passed to a remote Scheduler Server, which will trigger the Kepler Workflow Run Engine to execute the archival workflow to segment, document, and store the sensor data to Metacat. Users can search for and retrieve the archives in the Metacat server through the standard data search interface in Kepler.

Kepler also has a general workflow scheduler, which can be used to execute any workflow periodically. This allows sensor site administrators to automate the execution of, e.g., QA/QC workflows at an appropriate frequency.

B. Scientific Lifecycle (Sensor Data Usage)

To be useful within the scientific lifecycle of sensor data, a sensor management tool ideally provides powerful analysis, modeling, and visualization capabilities. Kepler provides hundreds of analysis and modeling functions, ranging from atomic signal and image processing functions to integration [3]. For example, to accomplish a quality assurance analysis within Kepler, one can use the sensor actor to feed a stream of data in real-time from a sensor, connect this to the R system to use R's excellent time series analysis tools to detect anomalies, and

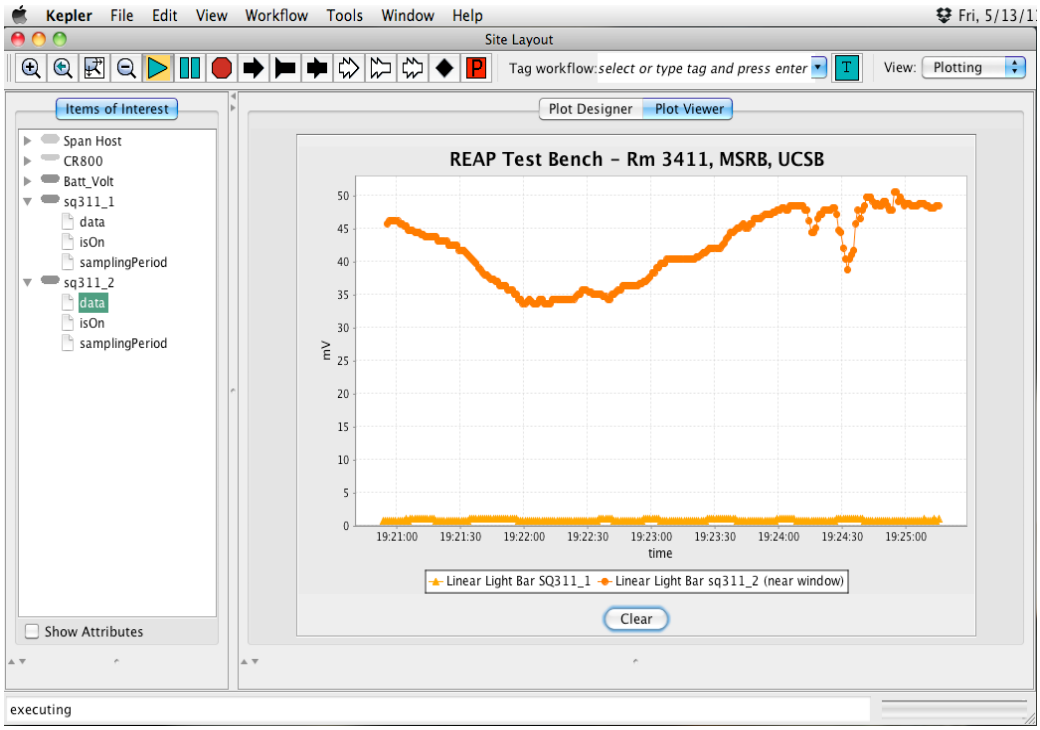


Figure 4. Kepler displays near real-time plots of sensor values over time, allowing scientists to quickly get a sense of the status of data collection at a site. When sensors are configured (e.g., to increase sampling rate), the results are immediately visible.

then feed the results of that analysis to Matlab for visualization. One could also annotate the data as the processing occurs, and output the resultant derived data set as a new data product, and save this in a data repository such as Metacat. In addition to real-time data access, Kepler provides tools to access the archived data sets from the engineering lifecycle, and a wide variety of other data from repositories around the world. Scientists can therefore combine data from historical periods with the data from real-time streams to detect changes in data trends over time and space. This flexibility allows scientists to mix and match the best analytical tools for the job while Kepler handles all of the orchestration and connectivity among the components. The end result is a workflow that fully documents the entire process that was employed to filter, transform, and analyze sensor data. In addition to these standard capabilities, we have added some additional features to Kepler specifically to help manage the scientific lifecycle of sensor data: real-time plotting, workflow scheduling for remote execution, and workflow run management.

Real-time plotting. The Kepler Engineering View provides the capability to plot live sensor data (Figure 4). The plotting view allows users to configure multiple plots and choose which sensors' data to display in each plot. A single plot can be configured to show the data from multiple sensors, allowing visual comparison of live data in near real-time. Useful plot interactions such as zooming, auto-range, labeling of title and axes, adjusting point shapes and colors, clearing data, and exporting to static image files are supported.

Scheduling and remote execution. The same scheduling subsystem that handles periodic data archiving workflows from the engineering lifecycle can be used in the scientific lifecycle to periodically run analyses and models as needed by the scientist. When scheduling a workflow to execute remotely, the scientist can choose the time period for the executions and the interval at which the workflow should be re-run. For sensor data that is being continuously generated, this is extremely useful to periodically produce statistical summaries, generate or update summary plots for display on websites, and run forecast and hindcast models.

Workflow run management. By allowing scheduled workflows to be run on remote servers, it can be difficult to track how many times a workflow has run, and for each run whether it succeeded or failed with particular error conditions. The Kepler Workflow Run Manager provides an interface to browse through all of the workflow runs that were executed on a local or remote instance of Kepler. A complete provenance record of each workflow run is recorded, and the Workflow Run Manager provides a graphical view of these runs. Runs can be tagged in order to cluster related runs together, and they can be searched based on the provenance metadata (e.g., to find all runs for a temperature data archiving workflow that were run after June 21, 2011). From the Workflow Run Manager, one can also open the workflow as it was when executed, view any reports that were generated, and save a workflow run from a local instance of Kepler to a remote repository for backup or to share with colleagues.

III. EVALUATION

A sensor simulator was created to simulate different types of sensor network deployments and to aid implementing and testing Engineering View components. The simulator provides a virtual sensor network. Configuration parameters include the number of sensors, sampling rates, and sensor metadata such as make, model, location, etc.

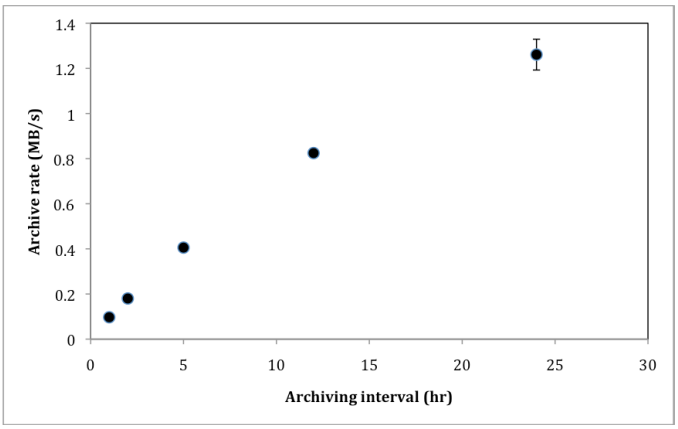


Figure 5. The archival rate as a function of archiving interval. The archival rate is the ratio of data size to the archiving workflow execution time.

To evaluate the performance of archiving sensor data, we measured the execution time of the archival workflow processing different amounts of data. We first ran the sensor simulator configured with one hundred sensors, each generating one sample per second (1 Hz). The simulator executed this configuration for 1, 2, 5, 12 and 24-hour periods, storing generated data into a DataTurbine server. We then ran the archival workflow to retrieve the samples from DataTurbine, and to create and upload the datasets into Metacat. The DataTurbine server and archival workflow ran on an iMac with a dual-core 3.2 GHz Intel CPU and 4 GB of memory. Metacat ran on an iMac with a dual-core 2.8 GHz Intel CPU and 2 GB of memory.

Figure 5 shows the archival rates for the different amounts of sensor data collected. The archival rate increases almost linearly with the sensor data. While the rate slows down for the 24-hour interval, we believe this shows good scalability when running the Kepler Sensor Platform server-side components on desktop hardware for the typical scale of sensor networks that we expect at the single laboratory level.

The archival workflow was executed three times for each archiving interval, and each point in the graph was calculated by averaging the execution times. For all intervals except for the 24-hour period, less than 3% standard error was observed. In the latter case, we believe the larger variability is due to occasional retries by the Metacat client uploading the dataset. The client attempts to transfer the entire dataset at once, and for larger sizes, retrying is more costly. We are planning to update the client to address this issue.

IV. Related Work

Although automated sensing is being widely used in environmental sciences, most of the existing technology focuses on data acquisition and analysis. However, as mentioned before, the sensor lifecycle also includes stages that relate to the engineering and health monitoring of a sensor network. There are commercial solutions targeting vendor-specific technologies and protocols, e.g., LabVIEW, LoggerNet and Simulink, which can be utilized for particular infrastructures. A drawback of commercial solutions is they are often too specific, hard to extend, or costly for small-scale scientific projects. In addition, there are open source initiatives like the Osiris and OOSTethys projects that are using the Sensor Web Enablement (SWE) specifications to build interoperable sensor webs. We will now discuss three commercial systems, followed by three open-source systems. To the best of our knowledge, our Kepler workflow-based solution is the only freely available open-source system that is vendor-independent, customizable, and extensible, allowing users to connect to, monitor, and control field-deployed hardware in an environment that also supports sophisticated statistical and modeling operations.

Campbell Scientific's LoggerNet [12] provides data access, monitoring and control for large datalogger networks. However, along with being a proprietary solution, LoggerNet does not allow complex analysis and models to be run on data values, and it does not allow automatic sensor network adaptation as a result of such analysis. In addition, LoggerNet does not document and archive data packages into a repository such as Metacat.

National Instruments LabVIEW [13] has a graphical user interface that integrates block diagrams with a dashboard interface. With its extensive hardware support that involves Field Programmable Gate Arrays (FPGAs), microprocessors and special purpose digital signal processors (DSPs), LabVIEW is a very versatile environment for custom data acquisition and analysis. However, LabVIEW is also proprietary software, and provides no archival capabilities.

Simulink [14] provides sensor platform support via the Simulink Coder (formerly Real-Time Workshop), targeting specific sensor hardware and architectures. Simulink Coder can generate embedded source code from Simulink diagrams and MATLAB scripts. Generated code can be used for real-time and other applications, including rapid development, simulation optimization, and testing with hardware in the loop. Although it provides some functionality needed for field-deployed hardware, the process for code generation and deployment can be cumbersome and requires specific target language compiler programs that are proprietary.

SensorKit [15] is an open source platform for sensor network management and data archival. As with the Kepler Sensor Platform, SensorKit uses SPAN to interface with dataloggers for data acquisition and device management. Data is archived to a SensorBase Database, and while this includes a web-based interface to graph, share, and export data, it does not provide sophisticated analysis and modeling capabilities.

The Viptos toolkit [16], derived from Visual Ptolemy, is an open-source system that is similar to Simulink Coder. Viptos models TinyOS-based wireless sensor networks via a graphical development and simulation system. TinyOS is an event-driven runtime environment used to build wireless sensor networks. Viptos builds on Ptolemy to enable the creation of flow diagrams, which are then used to create TinyOS programs from TinyOS components written in nesC, a programming language derived from C. Because Viptos is TinyOS-based, it has limited flexibility in sensor hardware choices. Nor does it provide documentation and archival features for sensor data streams.

The OSIRIS project [17] has developed a demonstration system for management of in-situ sensing data using the Sensor Web Enablement (SWE) suite of specifications from the Open Geospatial Consortium. For example, they developed web-based applications that use the Sensor Observation Service (SOS) [18] to access observations data from field-deployed sensors, and the Sensor Planning Service (SPS) for controlling sensors and sensor network components [19]. Similarly, the OOSTethys project has deployed SWE-based systems for managing sensors in ocean observing systems [20]. These and similar projects demonstrate the flexibility of OGC standards for accessing and controlling sensors, but do not provide graphical network visualization, data access, and control features within an application that supports sophisticated analysis and modeling. The union of standardized sensor network access with analysis and modeling tools as provided by Kepler would significantly strengthen these approaches.

V. CONCLUSION

We have described an extension to the Kepler scientific workflow system that supports full-lifecycle management of sensor networks. This extension addresses the needs of a wide audience, from technicians interested in monitoring and adjusting a site to keep it functioning effectively; to scientists that want to conduct complex analyses on sensor data streams, or compose workflows that will intelligently adapt a site's configuration in real-time in response to events of interest. The Kepler Sensor Platform supports scheduling QA/QC workflows to be run periodically on remote servers, provides an easy to use plotting view for quick comparisons of live data streams, and provides functionality for documenting, archiving, and retrieving sensor data into and from long term archives. Our tests have shown this extension effectively handles sites with many sensors, each sampling at a high frequency. Our work is entirely open-source, and thus may be utilized and extended by anyone with an interest. Future work will focus on interoperability with the Sensor Web Enablement suite of standards from the OGC.

ACKNOWLEDGMENT

This work was conducted as part of the REAP project with funding from the National Science Foundation (Grant # 0619060) and was supported by the National Center for Ecological Analysis and Synthesis (NCEAS), a Center funded by NSF (Grant #EF-0553768), the University of California, Santa Barbara, and the State of California.

REFERENCES

[1] B. Ludäscher, I. Altintas, S. Bowers, J. Cummings, T. Critchlow, E. Deelman, D. D. Roure, J. Freire, C. Goble, M. Jones, S. Klasky, T. McPhillips, N. Podhorszki, C. Silva, I. Taylor, and M. Vouk, “Scientific Process Automation and Workflow Management,” In Scientific Data Management: Challenges, Technology, and Deployment, Computational Science Series, A. Shoshani and D. Rotem, eds. chapter 13, 2009.

[2] Y. Gil, E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moraeay, and J. Myers, “Examining the challenges of scientific workflows”, Computer, vol. 40 (12), pp. 24-32, 2007.

[3] B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger-Frank, M. Jones, E. Lee, J. Tao, Y. Zhao, “Scientific Workflow Management and the Kepler System,” Special Issue: Workflow in Grid Systems, Concurrency and Computation: Practice & Experience, vol. 18, no. 10, pp. 1039-1065, 2006.

[4] D. Barseghian, I. Altintas, M.B. Jones, D. Crawl, N. Potter, J. Gallagher, P. Cornillon, M. Schildhauer, E. Borer, E.W. Seabloom, P.R. Hosseini, “Workflows and extensions to the Kepler scientific workflow system to support environmental sensor data access and analysis,” Ecological Informatics, vol. 5, pp. 3–8, 2010. doi:10.1016/j.ecoinf.2009.08.008

[5] S. Tilak, P. Hubbard, M. Miller, and T. Fountain, "The Ring Buffer Network Bus (RBNB) DataTurbine Streaming Data Middleware for Environmental Observing Systems", in Proc. eScience, 2007, pp.125-133.

[6] Ye W, Silva F, DeSchon A, Bhatt S. 2008. Architecture of a satellite-based sensor network for environment observation. In: NASA Earth Science Technology Conference (ESTC2008), Adelphi, MD, June 24–26, 2008. California, USA. University of Southern California, Information Sciences Institute: http://www.esto.nasa.gov/conferences/estc2008/papers/Ye_Wei_A8P2.pdf

[7] M.B. Jones, C. Berkley, J. Bojilova, M. Schildhauer, “Managing Scientific Metadata,” IEEE Internet Computing, vol. 5, no. 5, pp. 59-68, 2001.

[8] W. Michener, D. Vieglais, T. Vision, J. Kunze, P. Cruse, G. Janée, “DataONE: Data Observation Network for Earth — Preserving Data and Enabling Innovation in the Biological and Environmental Sciences,” D-Lib Magazine, vol. 17, no. 1/2, 2011, doi:10.1045/january2011-michener.

[9] E. Fegraus, S.J. Andelman, M.B. Jones and M. Schildhauer, ”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, 2005.

[10] A. Robin, S. Havens, S. Cox, J. Ricker, R. Lake, H. Niedzwiadek: OpenGIS sensor model language (SensorML) implementation specification. Technical report, Open Geospatial Consortium Inc. (2006)

[11] D. Higgins, C. Berkley, M.B. Jones, “Managing Heterogeneous Ecological Data using Morpho,” Proceedings of the 14th International Conference on Scientific and Statistical Database Management, July 24-26, 2002. J. Kennedy (ed). ISBN 0-7695-1632-7 ISSN 1099-3371.

[12] LoggerNet website. http://www.campbellsci.com/loggernet

[13] LabVIEW website. http://www.ni.com/labview

[14] Simulink website. http://www.mathworks.com/products/simulink

[15] F. Silva, A. Deschon, J. Chang, S. Westrich, Y.H. Cho, S. Gullapalli, T. Benzel, E.A. Graham, “SensorKit: A Flexible and Extensible System for In-Situ Data Acquisition,” in American Geophysical Union, Fall Meeting 2009.

[16] E. Cheong, E.A. Lee, and Y. Zhao. "Viptos: A Graphical Development and Simulation Environment for TinyOS-based Wireless Sensor Networks," Demo Abstract, Proceedings of the Third ACM Conference on Embedded Networked Sensor Systems (SenSys 2005), San Diego, California, USA, November 2-4, 2005.

[17] OSIRIS: Open architecture for Smart and Interoperable networks in Risk management based on In-situ Sensors. http://www.osiris-fp6.eu/

[18] G. McFerren, D. Hohls, G. Fleming, T. Sutton, "Evaluating Sensor Observation Service implementations," Geoscience and Remote Sensing Symposium, 2009 IEEE International, IGARSS 2009 , vol.5, pp.V-363-V-366, 12-17 July 2009. doi: 10.1109/IGARSS.2009.5417655

[19] S. Jirka, A. Bröring, C. Stasch, “Discovery Mechanisms for the Sensor Web,” Sensors, vol. 9, issue 4, pp. 2661-2681, 2009. doi:10.3390/s90402661

[20] L. Bermudez, E. Delory, T. O'Reilly, J. del Rio Fernandez, "Ocean observing systems demystified," OCEANS 2009, MTS/IEEE Biloxi - Marine Technology for Our Future: Global and Local Challenges , pp.1-7, 26-29 Oct. 2009.

The Atlas of Living Australia's Spatial Portal

Lee Belbin
Atlas of Living Australia
leebelbin@blatantfabrications.com

***Abstract*—The Atlas of Living Australia is an AUS$65m Australian Government initiative to "To develop an authoritative, freely accessible, distributed and federated biodiversity data management system". The Atlas, led by CSIRO, partners with over 18 National, State and Territory agencies to deliver online, a wide range of biological and environmental information. The Atlas also supports nodes of, or links to the Global Biodiversity Information Facility, Catalogue of Life, Encyclopedia of Life, Biodiversity Heritage Library (BHL), Map of Life, Barcode of Life Data Systems (BOLD), Ocean Biogeographic Information System, Morphbank, the Taxonomic Database Working Group and other projects. Two years into the three year project, the Atlas delivers over 114,000 species and 22 million occurrence records, 200+ environmental layers, a range of spatial and annotation tools and citizen science support. The Atlas Spatial Portal, the focus of this paper, is a tool designed to support environmental research and management. The focus of the portal is species, areas, environmental layers, spatial analysis and data import/export.**

Keywords—GIS; spatial; Australia; biodiversity; tools; analysis; environment; management; conservation.

I. INTRODUCTION

The Atlas of Living Australia [1] is a Federal Government initiative to provide public access to the widest range of biodiversity-related data in the Australian region. Funding for the project came from the national Educational Investment Fund (AUS$30m), the National Collaborative Research Infrastructure (AUS$8m) and from in-kind contributions from partner agencies (AUS$26.5m).

The mission of the Atlas is "To develop an authoritative, freely accessible, distributed and federated biodiversity data management system". Fortunately, at the start of the project, Australia had useful infrastructure in place for sharing biological data. The Commonwealth Heads of Faunal Collections and Commonwealth Heads of Herbaria provided an existing structure for the sharing of faunal data through the Online Zoological Collections of Australian Museums - OZCAM [2] and Australia's Virtual Herbarium [3]. To this base, significant volumes of data have been added from Birds Australia [4] and various State and Territory collections. The coverage of the Atlas includes plants, animals and microorganisms, marine, terrestrial and limnetic species, native and non-native species and endemic and invasive species.

There are two closely related projects: the Terrestrial Ecosystem Research Network [5] and the Integrated Marine Observing System – iMOS [6]. From the Atlas perspective, TERN focuses on systematic bio-survey data while iMOS has focused on monitoring the marine physical/chemical environment. In addition to infrastructure related to species occurrence records, the Atlas also has significant projects related to citizen science, the Bioversity Heritage Library - BHL [7], Barcode of Life Data - BOLD [8], Morphbank [9] and Identify Life [10] that link with Atlas data.

II. USER NEEDS ANALYSIS

The Atlas commissioned an extensive user needs analysis report [11]. The key applications identified were species distribution analysis, species identification, site assessment, habitat management and planning, managing reference databases, public education, synecology and biosecurity. Three issues of particular significance were identified: resolving scientific names; integrating amateur observations and the management of sensitive data. As well as a feedback link on all Atlas pages, a comprehensive annotations service for species data was also required.

The target audience of the Spatial Portal is the scientific community and the key statement in the analysis was "Distribution analysis is the dominant task. The ability to retrieve spatial information will be essential – varying in time, varying in scale, with many different forms of content." The spatial priorities distilled to "Where does this species occur?" and "What species occur in this area?" We have extended both functions to include any taxonomic level and 13 different ways of defining 'area'. We have also placed a high priority on the ability to upload and download data. Web service access to all key Atlas functions was considered as a basic requirement and feedback can be submitted from any web page of the Atlas.

III. THE SPATIAL PORTAL

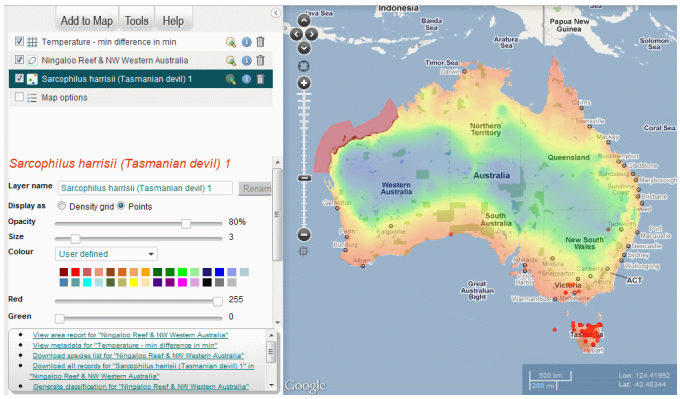


Figure 1. The Atlas of Living Australia's Spatial Portal.

The Spatial Portal of the Atlas (Fig. 1, http://spatial.ala.org.au) provides the main geocentric view of Atlas data. We have tried to align with Best Current Practice; building on what has been done effectively rather than reinvention. A review of over 30 existing geospatial portals provided useful criteria to address and a 'Google Maps/Earth look and feel' seemed to be assumed by most users.

To provide efficient functionality for the public and the research community was however no easy quest. We addressed this by trying to make simple and more complex functions equally intuitive. The 'Add to Map' tab provides rapid public display of taxa, areas and layers on the map while the Tools tab provide some form of analysis. Options for all functions are context sensitive.

We also attempt to lead the casual user to explore more advanced functions of the Spatial Portal. For example, the lower-left area of the Portal is used to provide hints of possible actions based on current mapped layers. For example, if a species has been mapped, the hint area provides links to species metadata, species records download, a scatterplot for the species and a spatial predictive model for the species. If an area is defined/mapped, hints include offering a checklist of all species in that area, a spatially predictive model of mapped taxa in that area or an environmental domains analysis for the area.

On the map window, there is only one function beyond the standard Google zoom, pan and 'zoom to my location functions': a layer interrogation button (a hover tool). On the command window, there are only three functions: 'Add to map'; 'Tools' and 'Help'. Taxa, areas and layers can be added to the map. Tools include species lists, sampling layer data at species locations, scatterplots, environmental domain generation and species spatial modelling.

Mapping data and most tool operations result in a map layer listed on the top left of the Spatial Portal. A set of icons for each layer provide on/off, layer type, zoom to layer extent, access to layer metadata and layer deletion. The center-left of the portal displays layer legends and some analysis results. The legend provides both the keys to the mapped layers and the ability to change the characteristics of the displayed layer. For example, points representing species locations can be sized, colored using a color palette or RGB slider bars, and the transparency adjusted. The legend also allows for colour faceting on various Darwin Core fields such as data provider, basis of the record and spatial uncertainty. This functionally will be extended to most Darwin Core fields.

The code base of the Atlas Spatial Portal came from the iMOS Ocean Portal [12]. Key components of the Spatial Portal include Java ZK code base, GeoServer, OpenLayers, GeoNetwork, Google API and OGC standards (predominantly WMS to date). The species occurrence data and their intersections with all the spatial layers are based on SOLR indexing of a Cassandra database. Most of the functions provided by Atlas components use RESTful JSON services. Atlas code and data are generally licensed under Creative Commons CC BY 3 [13].

A. Species

Taxa include point occurrence records (based largely on the Darwin Core standard [14], checklists (lists of species within a defined area) and 'expert distribution maps – polygons defining where a species should occur. The latter is a special case of the checklist. We hope to be able to include species tracking data in the next year.

An 'auto-complete' search strategy is used and scientific and common names with synonymies are supported. The auto-complete list also displays the type of record, taxonomic level and the number of occurrences.

Two additional taxa-related options are supported. A set of point coordinates or a set of Life Science Identifiers – LSIDs [15] can be imported in comma-separated variable (CSV) format for a portal session. The coordinate file currently supports three variables; a label, a longitude and a latitude. The option to import (CSV initially and then Darwin Core XML) up to 256 additional fields will be implemented in the near future and will support faceting of the records on all fields. The LSIDs can be of any taxonomic level entity held by the Atlas and can therefore be used to map and analyze assemblages. There is no QA performed on data uploaded for a session.

B. Areas

'Areas' correspond to objects held in our gazetteer, generated dynamically or uploaded to the Spatial Portal. The base used by the Atlas is an amalgamation of the 2010 Australian gazetteer [16], the Global Administrative Areas Database [17] and all of our named polygons (e.g., States and Territories) and classes (e.g., 'Forestry') from 'contextual' layers (see below). In all, there are 13 ways that an area can be defined by the Spatial Portal:

- Interact with the map (draw bounding box, polygon, point and radius or select area from contextual/polygonal layer);
- Search (radius centered on street address or a gazetteer polygon);
- Preset areas (Australia, world or current view);
- Upload (Shapefile, KML or WKT format) and
- Define environmental envelope.

The environmental envelope option is by far the most complex and powerful. This option uses slider bars on upper and lower bounds of one or more of the 150+ environmental layers to define an environmental combination and the corresponding area on the map. For example, you can identify where in Australia the mean annual temperature is between 10-12c and the precipitation of the driest quarter is between 150-250mm.

Sadly, the Australian Gazetteer only contains point locations. To enable users to determine what species are associated with a named gazetteer location, we have added an option to select a 1, 5, 10 or 20km radius around the points.

C. *Layers*

'Layers' are defined in a traditional GIS sense. In the Atlas, these can be terrestrial or marine and either 'environmental' or 'contextual'. Environmental layers are usually grids containing continuous values such as mean annual temperature. Contextual layers are usually polygonal in structure and contain class values. An example contextual layer would be 'Land Use' and a class within that layer could be 'Forestry'.

There is an obvious overlap between layers and areas. As noted above, we have included all the classes of contextual layers in our gazetteer even though some classes would be defined as named multiple polygons rather than single polygons. This strategy provides users with maximum flexibility in mapping features. For example, the polygons of the Land Use class 'Horticulture' can be mapped directly from the Areas option as well as the complete Land Use layer via the Layers option.

The 200+ layers available through the Spatial Portal (http://spatial.ala.org.au/layers) required two user-selection options. Layer selection can be done by an auto-complete with synonyms and tags supported. For example, precipitation and rainfall can be used synonymously. In some cases, we have included codes for well-known suits of layers. For example "Bio01" can be used as a shortcut for the "mean annual temperature" of the climatic layer suite of Hutchinson and Kesteven [18].

A two-level classification was also developed to guide new users through the layer library. The top level of this classification has the terms, area management, biodiversity, climate, distance, hydrology, marine, political, substrate, topography and vegetation.

IV. Spatial Portal Tools

Two workshops were run to determine what tools would be appropriate for the Atlas [19] and what data would be required to support these tools [20]. Criteria for evaluating tools included for example "accepted as best current practice", "wide applicability' and 'robustness'. Subsequently, the ability to import and export data was considered as a higher priority than adding novel tools. For example, technical users wanted to import species occurrence coordinates, append environmental and contextual values, export the records and then use the integrated data with their favorite desktop tools. A few exemplar tools were however required to demonstrate the wide range of applications that could leverage an extremely large volume of integrated biological, environmental and contextual data. We wanted to demonstrate a few of the possibilities.

As one of the reviewers correctly pointed out, the Spatial Portal is in part a "data discovery, integration and subsetting tool that produces customized subsets of data that scientists can use." Downloaded data comes with whatever (usually Darwin Core) attributes are supplied by the data provider, for example record identifiers. In most cases, LSIDs have been added for species if they do not exist in the original records. UUIDs have been generated for all defined areas and GIS layers.

Metadata for species data usually applies at the species collection level while metadata for 'GIS layers' is at the individual layer level. Analysis downloads also include a reference identifier that can be re-submitted to the Spatial Portal in subsequent sessions to re-create the analysis and the associated outputs and downloads.

All the tools can make use of any definition of area and in relevant cases, taxa. For example, after starting the Spatial Portal, the spatial prediction of taxa based on uploaded coordinates and available environmental layers over an area of choice (based on any one of the 13 options above) could be achieved in a few mouse clicks.

One of the highest priorities for the Atlas is addressing data quality or more appropriately the concept of 'fitness for use' [21]. For the biological data, this is the responsibility of the Data Management group within the Atlas of Living Australia project. The Atlas philosophy is to expose the data as received, enable annotations and only correct the 'bleeding obvious'. That said, the Atlas is in a good position to detect data issues and potential solutions and direct these to the data providers to address as needed. 'Fitness for use' would require a separate paper and is not addressed further here.

A. *Area Reports*

Like all other tools, the area can be predefined or generated on the fly using any of the 13 options above. The report lists the area (in square kilometers), the number of species, the number of occurrence records, the number of species polygons (from expert distributions and species checklists) and the number of related publications via BioStor [22]. From the report, the species lists and the list of full occurrence records can be downloaded as CSV-formatted files. The report also provides for direct mapping of all occurrences within the defined area.

B. *Species Lists*

A comprehensive list of all known species occurrences in the Atlas can be produced and downloaded as a CSV-formatted file for any defined area (a checklist); single or multiple polygons. The report lists the Family Name, Scientific Name, Common Name/s, Taxon rank, Life Science Identifier and Number of Occurrences. Filtering of sensitive species is performed according to the Atlas sensitive data service but no further analysis of the checklist is performed.

C. *Sampling*

This option is similar to the Species List except that all species occurrence records for any defined area are listed and downloadable in CSV format. Key Darwin Core fields [23] are included. Optionally, any combination of the 200+ layer values (environmental or contextual) can be appended to the occurrence records for download. This option also enables the appending of layer values to uploaded coordinates (which are treated as just another biological GIS layer), which can then be downloaded. Sampling is perceived as probably being the most useful tool for the key target audience (environmental scientists and managers). Data can be integrated, subset, downloaded and readily ingested into the tools of choice. Scripts are being written to simplify ingestion of the downloaded data into packages such as R [24].

D. *Scatterplots*

Checking for 'environmental outliers' among occurrence records was seen as one way of making effective use of the environmental layers to be found through the Spatial Portal. A scatterplot tool was however likely to bring a far wider range of benefits for examining the environmental conditions associated with species occurrences.

The scatterplot tool (Fig. 2) accepts any two taxa (a primary and a secondary) in any defined area and any two environmental layers. The scatterplot also identifies all possible environmental combinations within the defined geographic area. A grey-scale is used to display the spatial extent of the environmental combinations; darker cells represent small geographic areas while large areas are displayed as lighter cells.

The scatterplot tool is interactive: Selecting occurrence records on the scatterplot (environmental space) will select the occurrences on the map (geographic space). Occurrence records with missing values are also separately identifiable. Included and excluded subsets (with either including the missing value records) can be automatically generated from the scatterplot. Each subset creates a new mapped (bio-) layer within the Spatial Portal and can be used for any subsequent analysis and download.

In the near future, we will generalize the scatterplot tool to accommodate any pair-wise combination of environmental and contextual layers. For example, we see great benefit of being able to cross tabulate say land-use (contextual) and dynamic land cover (contextual) or mean annual temperature (environmental).

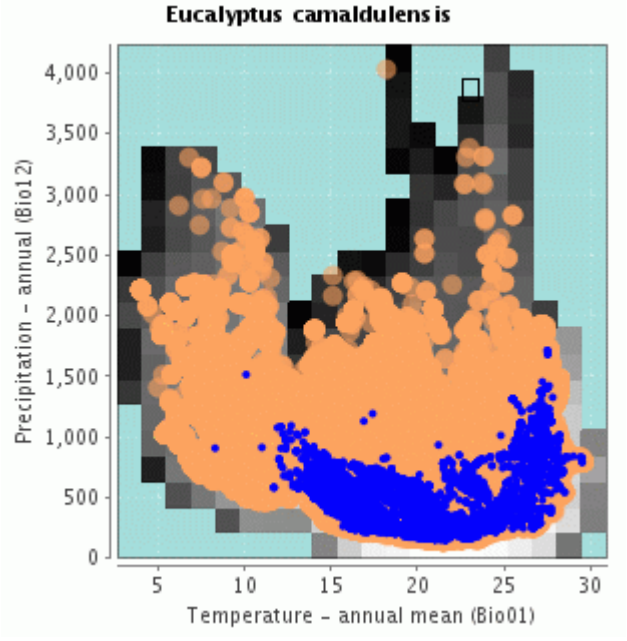


Figure 2. Scatterplot of *Eucalyptus camaldulensis* (primary – colored blue) with all *Eucalyptus* (secondary – brown) plotted against mean annual temperature and annual precipitation for contiental Australia. Greyscale envelope represents all possible terrestrial environmental combinations.

E. *Classification*

How do we make rational land management decisions where inadequate biological data is the norm? The use of environmental domains [25] may help. If species respond to environmental factors, it makes sense to use environment as a surrogate where adequate biodiversity data is lacking. Environmental domains result from a classification of multiple environmental layers.

The Spatial Portal contains a classification method [26] designed to generate environmental domains: areas of similar environmental properties based on multiple environmental layers (Fig. 3). One new environmental domains layer hopefully contains the most salient features of all submitted layers. The group colors are designed to represent group differences [27]. Fig. 3 provides a realistic ecosystem view of Australia based on only three environmental layers: annual precipitation; temperature annual minimum mean and a fertility scale based on lithology.

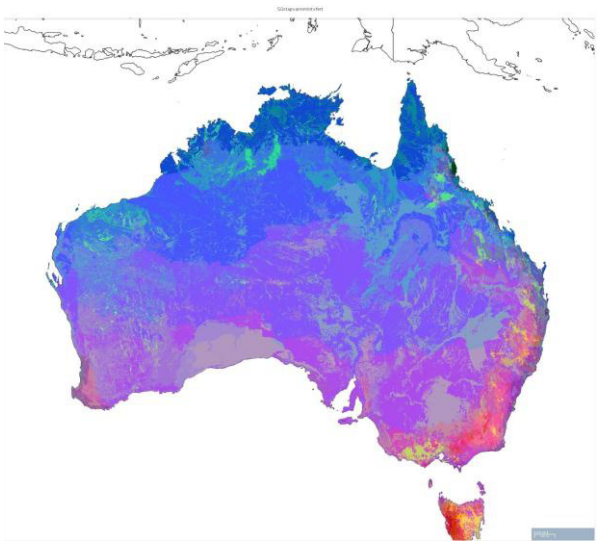


Figure 3. Spatial representation of the 51 group environmental domain classification of the Australian continent generated by the ALA Spatial Portal classification function using annual precipitation, temperature annual minimum mean and a fertility scale based on lithology.

F. *Prediction*

Spatial prediction models were seen as a stereotypic method demonstrating the value of integrated biological and environmental data. Such models provide environmental interpolation; displaying where species *could* occur and with what probability. Like any tool, it can be abused but that is not the focus of this paper.

MaxEnt (maximum entropy density estimation) [27] was recommended by the tools workshop [19] and is implemented in the Spatial Portal as an external package. Extensive testing suggested that realistic models could be generated from environmental layers that were known to constrain the distribution of the species in some way (Fig. 4).

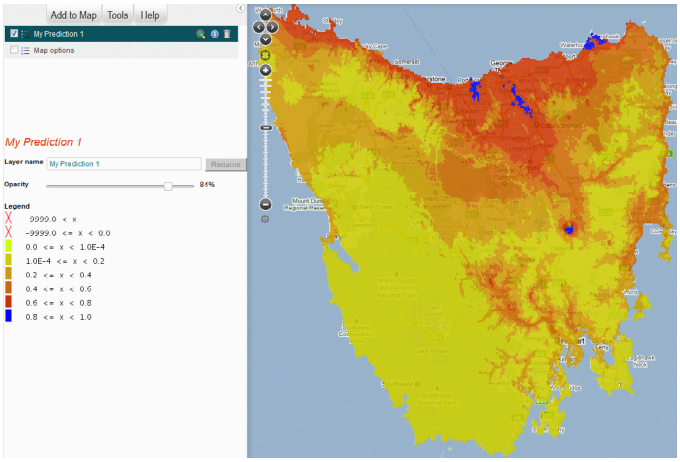


Figure 4. MaxEnt spatial prediction of Sarcophilus harrisii (Tasmanian Devil) using Aridity index - month max, Precipitation – seasonality and Temperature - annual max mean.

CONCLUSION

The Spatial Portal of the Atlas of Living Australia aims to address the user requirements of the research, environmental management, environmental consultant and environmental NGO communities for the provision of biological and environmental information. A free an open source (FOSS) strategy based on a Google Maps look and feel was foundational. Designing an interface that would support the casual user and the target audience was a challenge addressed by simplifying to key functions and context sensitive prompts.

The focus of the Portal is the integration of biological and environmental/contextual data. Integration of Australia's biological data is a significant end in itself, but linking these data with a wide range of environmental and contextual layers provides opportunities to support an extremely wide range of environmentally-related applications. Portal tools such as spatial modeling of species demonstrate the value of integration but the export (or import) of occurrence records with optional environmental and contextual values appended probably provides the greatest utility for the target audience. The current round of Government funding of the Atlas ends June 30, 2012. The project is based on open source principles and has established protocols with a wide range of organizations. This strategy aims at automating, minimizing and spreading the load of ongoing development and maintenance.

ACKNOWLEDGMENT

My appreciation goes to Donald Hobern, the Director of the Atlas of Living Australia for providing me with autonomy, a budget to achieve a significant outcome (that I leave others to judge), astute observations and endless encouragement. The Spatial Portal would not be possible without the enthusiasm, skill and dedication of the spatially scattered team: Adam Collins; Ajay Ranipeta; Angus MacAulay; Gavin Jackson and Brendan Ward. The comments of the three reviewers are also gratefully acknowledged. My thanks also go to Miles Nicholls and Bryn Kingsford whose tenacity in tracking down data and metadata will I hope be appreciated by many.

REFERENCES

[1] The Atlas of Living Australia: http://www.ala.org.au.

[2] Online Zoological Collections (OZCAM): http://www.ozcam.org.au/.

[3] The Australian Virtual Herbarium (AVH): (http://www.anbg.gov.au/chah/avh/avh.html.

[4] Birds Australia: http://www.birdsaustralia.com.au/.

[5] Terrestrial Ecosystem Research Network (TERN): http:/www.tern.org.au/.

[6] Integrated Marine Observing System (IMOS): http:/www.imos.org.au)

[7] The Biodiversity Heritage Library (BHL): http://www.biodiversitylibrary.org/.

[8] Barcode of Life Data (BOLD): http://www.boldsystems.org/.

[9] Morphbank: whttp://www.morphbank.net/.

[10] Identify Life: http:/www.identifylife.org/.

[11] The Atlas of Living Australia User Needs Analysis report http://www.ala.org.au/about/communications-centre/publications/user-needs-analysis-report/.

[12] iMOS Spatial Portal: http://imos.aodn.org.au/webportal/.

[13] Creative Commons CC BY 3.0 license: http://creativecommons.org/licenses/by/3.0/au/deed.en

[14] The TDWG Darwin Core standard: http://www.tdwg.org/activities/darwincore/.

[15] The TDWG Life Sciences Identifier Applicability Statement: http://www.tdwg.org/standards/150/.

[16] The Australian Gazetteer: https://www.ga.gov.au/products/servlet/controller?event=GEOCAT_DETAILS&catno=71110.

[17] Global Administrative Areas Database: http://www.gadm.org/.

[18] Hutchinson, M.F. and Kesteven, J.L., 1998. Monthly mean climatic surfaces for Australia. Unpublished. http://fennerschool.anu.edu.au/publications/software/creswww.pdf

[19] Flemons, P., Raymond, B., Brenton, P. and Belbin, L., 2010. Atlas of Living Australia report on the spatial analysis toolkit workshop. Unpubliched. 2010. http://www.ala.org.au/wp-content/uploads/2010/09/Lee_ALA-Spatial-Ananysis-Tools-Workshop-Report-V3.pdf

[20] Flemons, P. and Belbin, L., 2010. Report on the Environmental data library workshop. Unpublished. http://www.ala.org.au/wp-content/uploads/2010/08/Lee-NEDL_Workshop-Report-Final.pdf

[21] GBIF position paper on Fitness for use: http://www.gbif.org/communications/news-and-events/showsingle/article/gbif-position-paper-enhancing-fitness-for-use-across-gbif/.

[22] Page, R.D.M., 2011. Extracting scientific articles from a large digital archive: BioStor and the Biodiversity Heritage Library", BMC Bioinformatics, 12, 187. doi:10.1186/1471-2105-12-187.

[23] TDWG Darwin Core standard terms: http://rs.tdwg.org/dwc/terms/.

[24] The R Project for Statistical Computing: http://www.r-project.org/.

[25] Nix, H.A., 1986. A biogeographic analysis of Australian elapid snakes. In: R Longmore (Editor), Atlas of Elapid Snakes of Australia., pp. 4-15. Australian Flora and Fauna Series Number 7. Australian Government Publishing Service, Canberra.

[26] Belbin, L., 1987. The use of non-hierarchical allocation methods for clustering large sets of data. Australian Computer Journal 19, 32-41.

[27] Belbin, L. Marshall, C. and Faith, D.P., 1983. Representing relationships by automatic assignment of colour. The Australian Computing Journal 15, 160-163.

[28] Phillips, S.J., Dudík, M. and Schapire, R., 2004. A maximum entropy approach to species distribution modeling. In: Carla E. Brodley (Editor), Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Canada, pp. 655-66.

Rapid Development of a Hybrid Web Application for Synthesis Science of *Symbiodinium* with Google Apps

Erik C. Franklin, Michael Stat, Xavier Pochon, Hollie M. Putnam, Ruth D. Gates
Hawaii Institute of Marine Biology, School of Ocean and Earth Science and Technology, University of Hawaii at Manoa
erik.franklin@hawaii.edu, stat@hawaii.edu, pochon@hawaii.edu, hputnam@hawaii.edu, rgates@hawaii.edu

Abstract— **Data interoperability facilitates the integration, access, and delivery of information from a variety of sources to synthesize knowledge for scientific collaboration. Often the success of a workgroup-scale data integration project can be hindered by the insufficient computing expertise of the team, inadequate network resources, and limited funding to support cyberinfrastructure. We explore the utility of the free, cloud-based Google Apps to overcome these potential shortfalls and present a case study for the development of a hybrid web application, called GeoSymbio, that synthesizes global bioinformatic and ecoinformatic data of *Symbiodinium*, a group of uni-cellular, photosynthetic dinoflagellates that are found free-living or in symbiosis with a wide range of marine invertebrate hosts including scleractinian coral. Google Apps allowed our five member multidisciplinary group of biologists to develop a web-based tool to discover, explore, and visualize project data in a rapid, cost-effective, and engaging manner. Although the final product exceeded our expectations, there were certain limitations that we encountered including file data storage limits, the slow loading speed of some tools, and incomplete integration among applications. Traditionally, scientific data synthesis and integration has been presented as static journal review articles. Here, we demonstrate a path to develop a novel type of web-based, data-driven, and publically accessible review of scientific knowledge that allows the user to dynamically interact with the compiled information using Google Apps. GeoSymbio is located at https://sites.google.com/site/geosymbio/.**

Keywords—bioinformatics; data interoperability; ecoinformatics; GeoSymbio; Google Apps; Symbiodinium

I. INTRODUCTION

In March 2011, our group of five biologists was tasked with the compilation of global bioinformatic and ecoinformatic data on coral host-symbiont symbioses for analysis, synthesis and visualization as part of the "Tropical Coral Reefs of the Future" working group at the National Center for Ecological Analysis and Synthesis (NCEAS). Over the prior two years, we had already considered this issue and thus had created a data schema and populated a database with approximately 2500 records manually data-mined from GenBank and journal articles. Yet after the extensive early work on the project, the information only existed as a spreadsheet file circulated between our desktop computers. During the working group, we were challenged with the issue of how our small scientific research team could, in a rapid time frame for a low cost, integrate various data streams, expand the database, and then broadly share the synthesis results without having a computing support team (such as a network administrator, database administrator, or web programmer). Our proposed solution involved the adoption of the Google Apps software as the computing framework for data entry, management, and visualization of project information. By May 2011, we had a functional solution of web-based tools that exceeded our project requirements. In this article, we describe the development process relative to the compilation, integration, access, and delivery of information for the scientific synthesis of global symbiont data (of the genus *Symbiodinium*) with Google Apps.

II. BACKGROUND

A. *Symbiodinium Biology and Taxonomy*

The genus *Symbiodinium* is a group of uni-cellular, photosynthetic dinoflagellates found either free-living or in symbiosis with a wide range of marine invertebrates including scleractinian corals. *Symbiodinium* encompasses nine divergent genetic lineages called clades [1] which each contain multiple subclade sequence types. The Internal Transcribed Spacer 2 region (ITS2) of the nuclear ribosomal array has been used extensively for genetic identification and taxonomic description of over 400 distinct *Symbiodinium* subclade types in invertebrate hosts sampled from a variety of marine habitats of tropical and subtropical waters [2, 3, 4, 5].

B. *Global Symbiodinium Database Schema*

Prior to the NCEAS working group, we had previously designed a database plan to reflect the bioinformatic and ecoinformatic information relevant to global *Symbiodinium*-host symbioses (Table I). The plan had 33 variables that described information based on *Symbiodinium* occurrences such as sequence identification, method of identification, host taxa, collection event, sampling location and citation reference details. The variables and their definitions were adapted from the Ocean Biogeographic Information System (OBIS) Schema v1.1 [6] which is an extension of the Darwin Core Version 2 standard. The detailed definitions of each of the data fields are available online at the GeoSymbio schema webpage [7].

TABLE I. DATA FIELDS OF GLOBAL *SYMBIODINIUM* DATASET.

Group	*Field*	*Data Type*
Symbiodinium	Clade	Text
	Subclade	Text
	Gene	Text
	Isolate	Text
	Redundancy of Sequence	Text
	Species	Text
	Methodology	Text
	Genbank	Text
	Genbank link	Hyperlink
Host Taxa	Host Phylum	Text
	Host Class	Text
	Host Order	Text
	Host Family	Text
	Host Genus	Text
	Host Species	Text
	Host Scientific Name	Text
	Host AphiaID[a]	Text
	Environment	Text
Collection Event	Ocean	Text
	Country	Text
	State Region	Text
	Sub Region	Text
	Locale	Text
	Latitude	Numeric
	Longitude	Numeric
	Coordinate Precision	Numeric
	Minimum Depth	Numeric
	Maximum Depth	Numeric
	Start Year Collect	Date
	End Year Collect	Date
Citation	Reference short	Text
	Reference full	Text
	Reference link	Hyperlink

a. World Register of Marine Species unique taxonomic identifier (www.marinespecies.org)

C. Data-Mining GenBank and the Scientific Literature

The primary repository for *Symbiodinium* genetic sequence information is the US National Center for Biotechnology Information's GenBank. Sequence records are archived digitally, identified with an accession number, and accessible through a variety of online NCBI search tools. In 2009, we began querying GenBank for all *Symbiodinium* ITS2 sequences in order to populate the database. We quickly discovered that GenBank contained many redundant entries, records that were often incomplete, and that there was little quality control on the submitted ITS2 data. Furthermore, the missing or coarse resolution of geographic description often encountered in GenBank submissions severely limited our ability to automate the geographic mapping of genetic sequence data, an important requirement for our database. From the redundant sequences, we identified identical sequences (i.e., 100% residue similarity) with different accession numbers as synonyms with the first published record as the "parent" accession number. Then, we manually searched the source literature to confirm or ascertain the following descriptive characteristics for each sequence: host taxa, location, collection year, and laboratory methodology. The mapping of *Symbiodinium* occurrence locations often required reading the primary literature source identified in the GenBank accession record, with a cross-check of location in GEOnet Names and Google Earth. Although the process was time consuming, we had approximately 2500 records in our global *Symbiodinium* data table by March 2011.

III. DEVELOPMENT OF GEOSYMBIO

Building the capacity to examine the diversity, ecology and biogeography of *Symbiodinium*-host symbioses has global and societal implications and thus, the compilation and dissemination of this information was essential. One of the major barriers to progress was that the geographic, host taxa, and temporal details of the *Symbiodinium* occurrence records were not exposed and documented well in existing databases. This required manual examination of data records as well as extensive reading of the primary literature to extract useful ecological information to match with the genetic data. Our data-mining activities had already provided a good foundation for the dataset but we lacked a streamlined means to visualize and explore the data for research. To provide better access to this information, we determined that we required a system that provided four basic functions: (1) geospatial visualization, (2) text-based queries, (3) knowledge summaries, and (4) data products for further analyses. Given the time, personnel, and fiscal constraints, we required a simple, cost-effective (as in free), and robust solution for the system. After exploratory research of potential solutions, we began development of GeoSymbio using Google Apps in March 2011 at the NCEAS working group meeting.

A. Project Framework with Google Apps

Google Apps are a suite of cloud-based software that provides a variety of functionality for performing computing tasks. To meet our system functionality requirements, we utilized Google Sites to host the web application, Google Maps and Google Earth for geospatial visualization, Google Spreadsheets for data entry and management, Google Fusion Tables for data management and visualization, and Google Gadgets for data queries, knowledge summaries, and visualization (Fig. 1). Google APIs were also used to script minor components of the system to retrieve data from remote servers and share map data from Fusion Tables using Javascript, for example. Once the initial data was imported to a Spreadsheet, the project activities were primarily cloud-based.

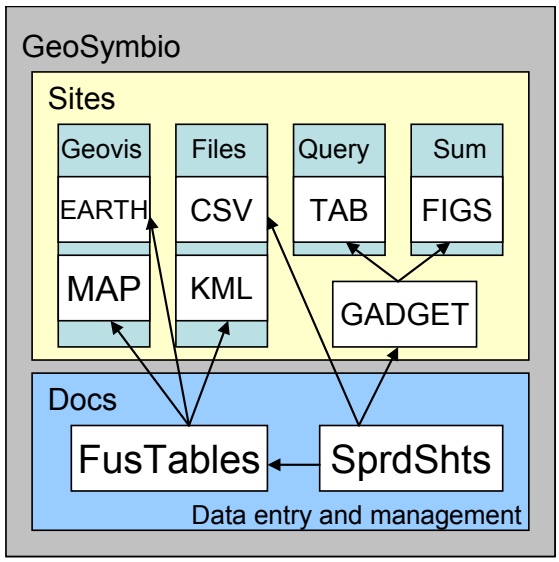


Figure 1. Schematic of the software components of the GeoSymbio web application. Components are white boxes, arrows are directional flow of data, and colored boxes are component groups based on function (green) or accessibility (yellow for public Google Site and blue for private Google Docs). Google Spreadsheets serves as the data entry point and primary management interface for the research team. Abbreviations in the figure are defined as Sites: Google Sites; Geovis: geovisualization, Files: files for download, Query: text-based tabular queries; Sum: knowledge summaries through dashboard figures; Earth: Google Earth; Map: Google Maps; CSV: a comma-separated tabular data file; KML: a keyhole markup language data file; TAB: table for text-based queries; FIGS: pie-chart figures of database element summaries; GADGET: Google Gadget; Docs: Google Docs; FusTables: Google Fusion Tables; and SprdShts: Google Spreadsheets.

B. *Data Entry and Management*

A Google Spreadsheets file provided the primary data entry and management interface for the *Symbiodinium* dataset. Previously, the dataset had been kept as a desktop spreadsheet file that was mailed to collaborators as new changes arose. This inefficient method of data management spawned multiple versions of the data file without a good means of tracking changes amongst the team. Furthermore, many additional *Symbiodinium* studies had been published which needed to be added to the database for the working group. Prior to the upload, we determined the most accurate version of the existing dataset for the project. Once in Spreadsheets, the data table allowed multiple simultaneous edits, versioning, and controlled vocabularies for data entry that greatly accelerated our ability to compile additional records in an efficient and robust manner. Several functions within Google Spreadsheets proved extremely useful for remote access of other data providers such as "ImportXML". For example, this function allowed genetic sequence retrieval from NCBI through their Entrez Programming Utilities (E-utilities) programs with XPath expressions. In addition, the RESTful structure of the applications allows direct access to the entire or subsets of the data files through the Google Fusion Tables SQL API. Using these methods, we nearly doubled the number of records to over 4800 and included records from all published studies of ITS2 gene. Once completed, the Spreadsheets data table was manually copied to a Google Fusion Table, an action that is not yet automated. The two data files in Spreadsheets and Fusion Tables provided the foundation for the other components of the hybrid web application, GeoSymbio.

C. *Hybrid Web Application*

GeoSymbio is the first comprehensive effort to collate and visualize *Symbiodinium* ecology, diversity, and biogeography in an online web application that is freely accessible and searchable by the public. The application structure is a hybrid or compilation that draws functionality and information from a variety of visualization tools and digital data and reference sources, with the core of the application hosted remotely or "in the cloud" using Google Sites [8]. The interconnected components of the application are made up of Google Spreadsheets, Google Fusion Tables, Google Maps, Google Earth, and Google Gadgets (Fig. 1). Thus, project information is accessible through any web-browser with internet access, so the application is not specific to a computer platform. The application is comprised of a collection of 10 web pages which include database knowledge summaries (DASHBOARD), searchable text-based queries (DATABASE) and spatial-based maps (MAPS and GOOGLE EARTH), the database schema (SCHEMA), a bibliography (BIBLIOGRAPHY), frequently asked questions (FAQ), downloadable map and sequence data files (DOWNLOADS), and project team contact information (CONTACT) (Fig. 2). The following sections of the paper

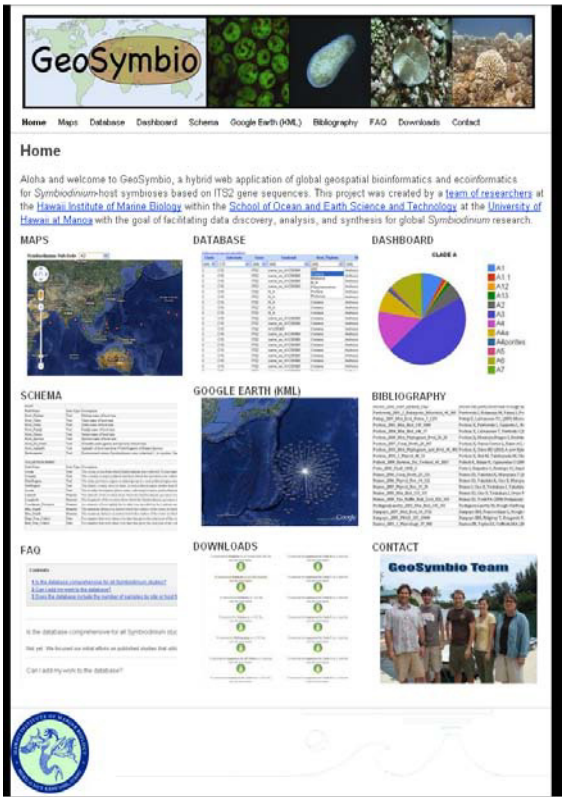


Figure 2. Screenshot of GeoSymbio hybrid web application home page.

detail the site functionality of geospatial visualization, text-based queries, knowledge summaries, and data downloads. The GeoSymbio URL is https://sites.google.com/site/geosymbio/.

D. Geospatial Visualization

Within GeoSymbio, the maps and Google Earth pages provide geospatial searches and visualization of the dataset for *Symbiodinium* clades and subclade types. The data for both mapping methods are stored in a Google Fusion table. The map components access the data through AJAX using the unique numeric identifier associated with the Fusion Table. Building off of the basic tutorials for Fusion Tables, we customized the map page interface with Javascript to allow a user to select the clade or subclade with buttons or a drop-down menu, respectively. The KML data network link for Google Earth is a standard feature available for Fusion Tables and did not require customization. The network link was used for both the Google Earth embedded viewer in the web page and the creation of the KML download file. The GeoSymbio maps webpage allows searches for *Symbiodinium* clade and subclade type as determined by ITS2 sequence type (Fig. 3). The Google Earth (KML) page provides a dynamic globe embedded in the website with the attributes of the GeoSymbio database accessible for each location in pop-up info windows.

E. Text-Based Data Queries

The GeoSymbio database page provides a dynamic data table with text filtering and grouping functions, which provide extremely flexible means to query for data. This functionality is provided by a Google Table Gadget that draws data from the primary data table in Spreadsheets. Filtering the database allows a simple yet powerful method to examine combinations of single filters for each attribute column. For example, a researcher interested in the occurrence of a subclade type within a particular host could filter dynamically to view records that meet the criteria. The grouping method of the database lends an even greater capacity to summarizing data with hierarchical relationships among the database attributes. To continue the previous example, a hierarchical grouping of host and clade with a count by subclade would dynamically update the table to show the selected criteria with subtotal record counts by group elements.

F. Knowledge Summaries

The knowledge summaries represent a quick view of the information contained in the dataset. The dashboard page presents a set of interactive pie charts that visualize the number and proportion of data records by *Symbiodinium* clade, ITS2 subclade sequence type, taxonomic order of the host, collection year, and location. These pie charts are Google Gadgets that pull data from the summarized subsets of information in the dataset. The visual nature of the charts presents a powerful means to rapidly convey relationships between fields in the dataset. The charts are dynamically updated as data is added to the Spreadsheets dataset. The charts can also be dynamically queried for the count and proportion of records for each category. In addition, the database schema and bibliography are both dynamically linked to the database and displayed as embedded table Google Gadgets on their respective webpages.

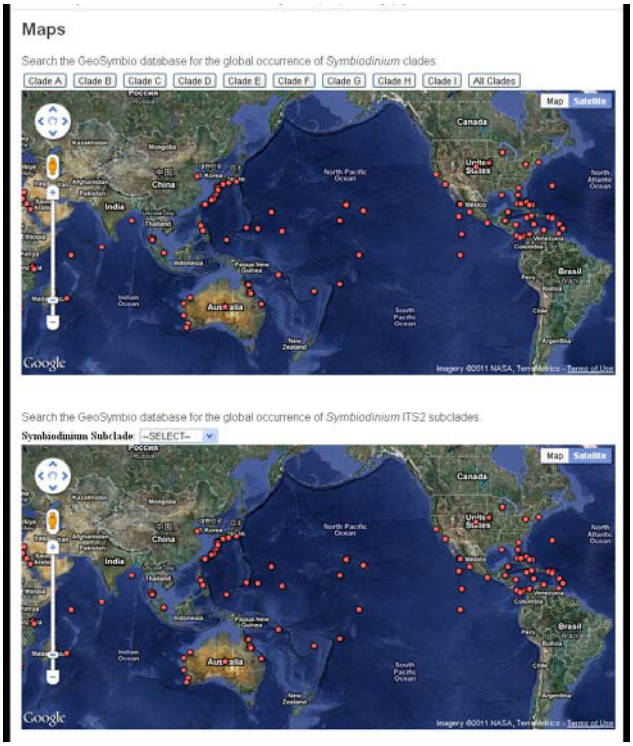


Figure 3. Screenshot of GeoSymbio maps webpage for geographic visualization of *Symbiodinium* clades (upper map) and ITS2 subclade sequence types (lower map) in embedded Google Maps and data from Google Fusion Tables. Search critieria can be subset by user in both maps with buttons and menus.

G. Data Downloads

The data download page includes links to download the map files (as .kml and .shp) to view the data in map programs such as Google Earth or ESRI ArcGIS. In addition, a set of genetic sequence alignment files (.fasta) can also be downloaded. Each of the nine sequence alignment files (including all sequences from one of each of nine existing *Symbiodinium* clades) was subjected to the following three steps. First, the sequence alignment file was imported into the alignment software BioEdit v7.0.9 [8] where it was subjected to automatic alignment using ClustalW [9] and further improved manually. Second, the aligned sequence file was run in 'DNA to haplotype collapser and converter' freely available at the online FaBox [10]. Except for the sequence and shapefile files, all other files were created through Google tools linked to the dataset.

H. Limitations of Google Apps

The overall development process with Google Apps was a strong success since we met our project requirements in a rapid and cost-effective manner. Nonetheless, there were elements about Google Apps that were not ideal including data storage limits, slow page loads, and less than seamless integration between applications. For example, the file data storage limit for a table is currently 400,000 cells. With the 33 variables in our data schema, we will be limited to approximately 12,000

records in the data table. At that point, we would possibly need to reconfigure the structure to multiple tables. Fusion Tables offers more data storage but not the same data editing and management functionality as Spreadsheets. Further, there is no current way to automate the association between a Spreadsheet and a Fusion Table which necessitates a manual update between the two. Also the slow load speed of several of the Google Gadgets and, in particular, the embedded Google Earth viewer requires 30 seconds to several minutes of wait time. These load times seem to improve after the first page loaded but may detract the casual user from using the tools by clicking away from the page during the delay. Furthermore, the high rate of change of infrastructure and functionality of Google Apps represents both an advantage and a disadvantage for this type of web solution. The advantage is that desired features may be implemented much more quickly than in a commercial off-the-shelf package, but the disadvantage is that the way things work can change without notice. Optimal performance was noted with the following browsers: Google Chrome v10, Microsoft Internet Explorer v8, and Apple Safari v4. These concerns in sum suggest that Google Apps may be optimal for smaller datasets and workgroup or smaller project teams. It is unclear if the free tools are scalable for larger projects.

IV. CONCLUSIONS

The need for a tool like GeoSymbio arises from the difficulties of integrating multiple data sources and information to perform bioinformatic and ecoinformatic data synthesis particularly in a geospatial context. These tasks can be challenging to execute without an interdisciplinary skill set of highly specialized scientific knowledge and a strong computing background, thus creating a barrier for progress among researchers. We demonstrate the rapid, cost-effective, and successful implementation of a hybrid web application for synthesis science developed with Google Apps. The web application provides four primary functions: (1) geospatial visualization, (2) text-based queries, (3) knowledge summary, and (4) data products. Starting with an existing data schema, the web application was developed and fully functional over a 5-week period from March to May 2011. Some disadvantages of using Google Apps include file data storage limits, the slow loading speed of some tools, and incomplete integration among applications. The rapid pace of development of Google Apps presents the benefit of an expanding suite of functionality but the potential for unwanted change with limited notice. Although we have expressed some caveats regarding the tools, we strongly endorse Google Apps for workgroup-scale projects that seek interoperability between various datasets and a set of web-based tools for dynamic exploration, synthesis, and sharing of knowledge on a scientific topic.

ACKNOWLEDGMENT

The National Marine Sanctuary Program (memorandum of agreement 2005-008/66882), the US Environmental Protection Agency Science To Achieve Results (STAR) PhD Fellowships (FP917096 and FP917199), the US National Science Foundation (NSF) grants through Biological Oceanography (OCE-0752604, OCE-1041673) and the Long Term Ecological Research (LTER) program (NSF 04-17412) provided financial support for this research. This work was conducted as a part of the "Tropical coral reefs of the future: Modeling ecological outcomes from the analyses of current and historical trends" Working Group supported by the National Center for Ecological Analysis and Synthesis, a Center funded by NSF (Grant #EF-0553768), the University of California, Santa Barbara, and the State of California. This is HIMB contribution # 1460 and SOEST contribution #8370.

REFERENCES

[1] Pochon, X., and Gates, R.D., 2010. A new *Symbiodinium* clade (Dinophyceae) from soritid foraminifera in Hawaii. Molecular Phylogenetics and Evolution 56:492-497.

[2] Lajeunesse, T.C., 2005. "Species" radiations of symbiotic dinoflagellates in the Atlantic and Indo-Pacific since the Miocene-Pliocene transition. Molecular Biology and Evolution 22: 570-581.

[3] Stat, M., Carter, D., and Hoegh-Guldberg, O., 2006. The evolutionary history of *Symbiodinium* and scleractinian hosts—Symbiosis, diversity, and the effect of climate change. Perspectives in Plant Ecology, Evolution and Systematics 8:23-43.

[4] Correa, A.M.S., and Baker, A.C., 2009. Understanding diversity in coral-algal symbiosis: a cluster-based approach to interpreting fine-scale genetic variation in the genus *Symbiodinium*. Coral Reefs 28:81-93.

[5] Silverstein, R.N., Correa, A.M.S., LaJeunesse, T.C., and Baker, A.C., 2011. Novel algal symbiont (*Symbiodinium* spp.) diversity in reef corals of Western Australia. Marine Ecology Progress Series 422:63-75.

[6] Vanden Berghe, E., 2007. The Ocean Biogeographic Information System: web pages. Available on http://www.iobis.org. Consulted on [6 June 2011].

[7] Franklin, E.C., Stat, M., Pochon, X., Putnam, H.M., and Gates, R.D., 2011. GeoSymbio database schema webpage. Accessed 28 July 2011 <https://sites.google.com/site/geosymbio/schema>

[8] Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41: 95-98

[9] Thompson, J.D., Higgins, D.G., and Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22:4673-4680.

[10] Villesen, P., 2007. FaBox: an online toolbox for FASTA sequences. Molecular Ecology Resources 7:965-968.

Searching for satellite data sets using Kepler, Metacat and EML

James Gallagher[1], Ben Leinfelder[2], Nathan Potter[1], Derik Barseghian[2]
1 OPeNDAP
2 NCEAS, University of California Santa Barbara
jgallagher@opendap.org, leinfelder@nceas.ucsb.edu, npotter@opendap.org, barseghian@nceas.ucsb.edu

***Abstract*—We present the results from building a data query module within the Kepler scientific workflow application. Our work focused on the query component of a larger use case from the Realtime Environment for Analytical Processing project where satellite-derived Sea Surface Temperature data were used to build match-up data sets as part of a workflow process. Kepler's integrated query capabilities allowed us to locate data described using the Ecological Metadata Language specification that was housed in a Metacat data catalog. Satellite data sets are significantly different from more traditional ecological data typically stored in Metacat, and while the resulting system worked well, it also highlighted areas where both Metacat and other satellite data-server software could be improved.**

Keywords—searching; workflow; integration; satellite data.

I. INTRODUCTION

Our motivation for this work was to determine how well suited the Kepler, Metacat and Ecological Metadata Language (EML) software components were for storing and querying descriptions of physical oceanography data. These data sets are often structurally very different from other traditional ecological data sets, although Sea Surface Temperature (SST) values are *fundamentally* ecological data. Kepler and Metacat have successfully provided effective storage and querying capabilities for ecological data, but we were curious to see how adaptable the spatial and temporal storage and search facilities provided by Metacat and EML would be to satellite data.

We found that many of the issues encountered were primarily due to the difficulty of building generic search systems for satellite data sets. The often-heterogeneous data storage schemes that are utilized by different data providers indicate a need for the server interface to abstract and generalize the details of any particular storage technique. While this can be accomplished using existing data servers for certain types of satellite data, to do so in the general case is difficult.

Providing effective query systems for satellite-derived data is important because these data sets are likely to become both more voluminous and more numerous. The 2007 National Academy of Sciences report on Environmental Data Management at NOAA estimated that satellite data volumes at NOAA alone will grow from ~3.5PB in 2007 to a projected level of over 40PB in 2020. It is very likely that current data organization patterns will persist and data discovery systems will face challenges similar to those documented here, but at a significantly larger scale. In fact, data management activities associated with storing and providing access to these data is considered "a significant data management challenge." [1]

A. The Relationship between Kepler, Metacat and EML

Kepler is open-source software that allows users to create scientific workflows, which are formal representations of the processes involved in scientific analyses. Kepler workflows may be used to connect a range of disparate software, and are saved in formats that are easily exchanged, re-run, versioned, and archived. [2] Metacat provides data set cataloging services to Kepler using Ecological Metadata Language (EML) documents. The EML specification [3] includes four basic element types; the *dataset* element is used to describe "broad information such as the title, abstract, keywords, contacts, maintenance history, purpose, and distribution of the data themselves." EML also contains elements for specifying spatial and temporal coverage of the data and supports grouping of related data objects into a single aggregated data set.

Metacat is designed to store, index and query any XML document, but is tailored for dealing with EML. Kepler's search system is able to quickly find EML-described data using predefined sets of indexed fields, including the use of temporal and spatial constraints. The Kepler workflow system interacts with Metacat using EarthGrid (formerly "EcoGrid") web services (see Fig. 1). The EarthGrid connects a number of independent systems and networks, providing access to data and metadata stored at distributed nodes. [4]

It would seem that given the Metacat and EML features for spatial and temporal data, they would be well suited for storing and querying metadata that describes satellite-derived data sets. However, this repurposing posed new development challenges.

B. The REAP Project and the Ocean use-case

The Realtime Environment for Analytical Processing (REAP) project consists (in part) of two very different use-cases which both use the Kepler scientific workflow system. These use-cases were specifically chosen to highlight fundamental assumptions inherent in the design of Kepler and to explore different solutions to the issues presented by these use-cases.

As different solutions were examined, we chose those that both directly addressed the needs of the specific use-case and that also could be reused more generally. In this paper we will discuss implementing search features for the *Ocean use-case*.

The Ocean use-case entails building match-up[1] data sets for comparing different Sea Surface Temperature (SST) data sets. While a description of the complete use-case is beyond the scope of this paper (see [5] and [6]), it is important to note that the data sets in this use-case are very different from those used in a typical ecology scenario - the subject of REAP's *Terrestrial Ecology* use-case [5]. The SST data sets used by the Ocean use-case are composed of thousands of individual files (e.g., a single Group for High-Resolution Sea Surface Temperature (GHRSST) data set at NOAA contains on the order of 10,000 files), each one holding SST values from one pass of a satellite. Roughly speaking, SST data sets may contain information in typical cartographic map projections (e.g., Lambert conformal) or they may contain data in *satellite coordinates* (i.e., each scan line is another increment along the satellite ground track) [7]. In this work we focused solely on those SST data sets that used a cartographic map projection where each pass contains data for the same geographic location. For all of the data sets used here, each file contains data corresponding to one satellite pass in the loose sense that while each scan line is actually a separate temporal event, they are clearly distinct from the scan lines captured by other passes of the satellite over the same geographic area. All the data files considered here are stored in 'self-describing' formats such as HDF4 or NetCDF that contain both data and metadata. Collections of these files are typically grouped into aggregate data sets, even though each file can also be considered a data set.

To satisfy the use-case, a user must be able to search for SST data sets that match spatial, temporal, and resolution parameters. Matches can then be input into the match-up data set workflow in Kepler. Difficulty arose when we attempted to implement this solution using the existing tools available within Kepler. As we will show, the issues are not specific to the implementation of Kelper but instead arise from fundamental data organization and representation paradigms used by our chosen data cataloging and search system.

C. *A bit more about the workflow*

In the Ocean use-case workflow, a user searches for a suitable SST data set to feed into the processing pipeline. The user must be able to search for data sets that intersect a region of interest specified by latitude, longitude, and time. In addition they must be able to narrow the search using both image resolution and parameter type. From a list of candidate data

[1] The term *match-up* refers to data sets that provide the same measured parameter for the same geospatial location (or, more generally, set of locations) using different sensors. For example, a match-up data set might consist of SST values from satellite data and SST in situ measurements from bouys.

Figure 1. Data flow from the OPeNDAP server to Metacat/EarthGrid and Kepler. This is the data flow for information from the servers to the search system to the search client. OPeNDAP servers provide a hierarchy of THREDDS catalogs that describe all of the granules served. The THREDDS catalogs are crawled and the resulting granules are aggregated, resulting in EML documents. The EML documents are then stored in MetaCat which provides an API that Kepler uses to perform searches that return EML documents.

sets, one is chosen and used as input to the data processing pipeline. The data processing software is composed of a set of legacy software, written in Fortran, that reads data from a sequence of satellite images.

The data sets targeted by this use-case are large (greater than 10 GB) and are stored at a variety of government and university research laboratories where they are accessible using remote data servers. Because the data are staged at many different remote locations, it is important to be able search for them through a unified catalog system.

The SST data in this use-case are accessed using servers that implement the Data Access Protocol (DAP) developed as part of the Distributed Oceanographic Data System (DODS) and now extended and maintained by OPeNDAP [8]. The DAP provides a way to access remote data over HTTP and enables clients to request subsets of data using a *constraint expression* [8]. Using constraint expressions can both reduce data transfer sizes and relieve clients of performing subsetting tasks. Most DAP servers are used with data sets that are stored in files or groups of files (as is the situation for this use-case) and provide a discrete URL for each file. The URL is used to access the data in the file; each access can be made using a constraint expression; and each URL can provide metadata about the data contained in the file. DAP servers provide an additional service to clients: they shield them from having to know about the actual storage format of the data. The servers translate the data into the DAP data model for transport, so all data is sent over the network using the same representation regardless of the data set's native storage format.

II. THE SEARCH INTERFACE

The search interface was implemented as an *actor* in the Kepler workflow system. Kepler uses the term *actor* to denote any workflow component and to separate the components of a workflow from the overall workflow orchestration. There are many different kinds of actors including ones aimed particularly at scientific applications: remote data and metadata access, data transformations, data analysis, interfacing with legacy applications, Web service invocation and deployment, and provenance tracking. Once dragged to the workspace, the Advanced Search actor we developed automatically displays a dialog used to enter the search criteria (see Fig. 2). When the user clicks *Search* the dialog will be replaced with a list of data sets that match the specified criteria (see Fig. 3) and the user may choose one. The output of the actor is a list of URLs. This list of result URLs is then routed to the processing software as part of the workflow (examples of workflows can be seen in [5]).

The searching system relies on EML records stored in Metacat and accessed using the EarthGrid web services (see Fig. 1). The EML records are built and inserted into Metacat using a data server crawler that reads metadata from a predefined set of DAP servers and, using a simple rule-based system, builds EML documents describing the data sets it finds. A complete description of this crawler/aggregator software is beyond the scope of this paper but one important feature is that it identifies common patterns of multi-file satellite data sets and builds aggregations for them using EML. It does this by examining large collections of URLs collected from a site and grouping subsets of those URLs using patterns. Thus groupings (i.e., aggregations) of the URLs can be formed without the data provider making them explicit. The aggregator component of the software then encodes these aggregations using EML so that its output easily integrates into the Metacat/EarthGrid/Kepler system.

The structure of the EML records used by the query system is shown in Fig. 4. As discussed previously, the EML *dataset* element holds information about the aggregation, while information about each file that makes up the aggregation is held in an *otherEntity* element. In Fig. 4 the structure of the *physical* child element of *otherEntity* is shown. This is the element actually used to bind the URL that references a single file with a specific date and time.

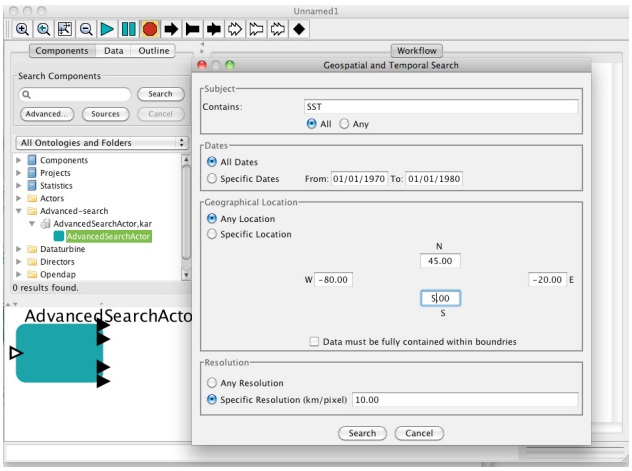


Figure 2. The search system interface implemented as an actor in the Kepler workflow system. Results from the search can be reviewed in a second pane and then fed into subsequent stages of a workflow (not shown).

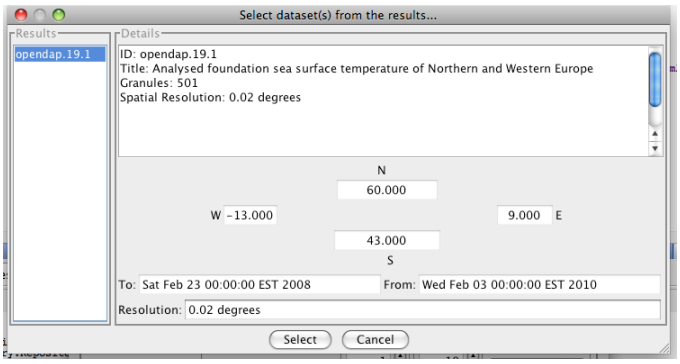


Figure 3. The result dialog. Users choose a single data set from the list of matches (here only one match was returned–shown in the left pane), browse metadata matching the query parameters and feed the result into the workflow.

```
<dataset/>
        <title/>
        <otherEntity/>
                <physical/>
                        <objectName/>
                        <encodingMethod/>
                        <url/>
                        <coverage/>
                                <temporal/>
                                <spatial/>
        ... (more otherEntity elements)
```

Figure 4. Within the *otherEntity* element, information about a single file is held in a *physical* element that contains a number of other child elements. This is repeated for each file in the data set. Spatial and temporal information is held in the *coverage* element.

III. DISCUSSION

One of the most important issues we confronted when building the search actor was how to handle the multi-file nature of the satellite data sets that play a central role in the Ocean use-case. Below we review the three approaches we considered and compare their merits and weaknesses. We found that while Metacat was generally flexible and extensible, there were

certain features of the system that required us to implement additional client side query refinements.

A. *Building aggregations for multi-file satellite imagery*

We examined three ways to form aggregations of the satellite data sets in this use-case:

- Building aggregations using the search system
- Building aggregations using the data servers
- Extending Metacat to better support aggregations

B. *Building aggregations using the search system*

Using EML documents to hold the aggregations provides a solution with a number of trade-offs. It de-couples the grouping of data URLs from the server, so the search system is no longer dependent on data providers building server-side aggregations. In this particular use-case, even though the technology for building and serving aggregations was available, it was not installed for most of the data sets. Even if the technology is installed, it may not be fully used, particularly by smaller laboratories, since it does require effort to configure and maintain. Moving control of the aggregations to the search system makes it easier to ensure uniformity among the data sets retrieved.

Building aggregations within the search system, however, has a number of drawbacks. First, the distributed nature of the system is subverted in that content generation cannot be spread among many different people and organizations. Building aggregations in the search system requires curation and aggregation be performed by the maintainers of that system. The complexity of the data sets compounds the problem; local experts may have knowledge about the data that the maintainers of the search system do not. There may be a considerable qualitative advantage to having the data provider build an aggregation. Another drawback is that some of the data abstraction capabilities a data server typically provides are lost. By building EML records that explicitly enumerate each URL in an aggregated data set, we are encumbering the client (the search interface in this case) with the task of selecting which of those URLs satisfies the search criteria.

An alternative approach to using single EML documents to hold the aggregations is to have Metacat store a single EML document for each URL in every data set. Using this scheme, the search interface would return all of the URLs that match the search criteria and the search interface would be responsible for forming the aggregations. If the aggregation step were skipped, one might assume that all returned imagery perfectly matched the search criteria. But a database with records for many SST data sets will likely return mixed records from different data sets that share the same space, time, and parameter values with similar resolutions but, for example, that differ in the specific algorithms used to compute the SST values. The user of such a system would be left to sort through tens or hundreds of thousands of URLs – effectively they would have to form the groups 'by hand,' an almost impossible task given the number of discrete items involved. Thus the search criteria used by the interface are necessary but not sufficient to select specific URLs for input in to the workflow in the general case.

We still could have adopted the *one EML document for each URL* scheme by building more intelligence into the search interface itself. When the interface received what would likely be 10,000 or more EML documents as the result of its query, it could have grouped those using other metadata in the documents. For example, it could have used the data set[2] title and the host name in the URL to form groups that would likely be correct.[3] However, doing this presents no real advantage over the case where a single EML document stores all of the URLs. The search client still must understand that the results of a query should be grouped before they can be used (so information-hiding is lost) and the task of forming the aggregations is moved away from the data sources to the search system (distribution of responsibility is lost). Building the aggregation capability into the search interface has one additional drawback. If the software that forms those groupings is found to have a flaw, it will have to be fixed in a subsequent release. However, a flaw in the EML stored in the Metacat database can be fixed by editing the EML document.

C. *Building aggregations using the data servers*

Using data server aggregations, a collection of two-dimensional 'granules' where the granules vary only in time can be combined to form a single three-dimensional data set. The DAP subsetting feature can then be used to access a latitude/longitude/time subset from this larger three-dimensional data set. That is, the data access operation performs the temporal search and subsetting operations. Effectively, that part of the search problem has been factored out of the search system and moved into the software that reads the data.

Building aggregations using the data servers is a technique that presents several significant advantages. The search system can store compact records that describe each aggregated data set, eliminating the coupling between the internal organization of the data sets and the search system. At the same time, this approach frees the search system's database maintainer from having to form the aggregations. Users of the workflow can be confident that the aggregations represented by the system are valid because the people closest to, and knowledgeable about, the data have built them.

---

[2] Note that when using DAP servers; each file is considered a unique data set. The aggregations are logical groupings that are imposed on the discrete elements. Forming an aggregation using a DAP server does not mean the individual URLs are not also accessible.

[3] Another solution is to include the equivalent of a foreign key in the EML documents so that the search interface can know which are related. This is really only an incremental improvement, however, since the search client still has to know how to use the key (information hiding is lost).

Unfortunately support for server-side aggregation is far from universal, and, in fact, it could not be applied to any of the SST data sets used by this use-case. Even if server-side aggregations were available for the data sets in this specific use-case, relying on them would violate our goal of generality. Because we assumed, for the sake of simplicity, that the use-case would handle only data sets with uniform cartographic projections covering the same geographic area, we eliminated a significant number of potential data sources that only provided data with satellite coordinates. We would like to use data stored in satellite coordinates [7], but the individual files in these data sets cannot be aggregated using the simple techniques applied to cartographically and geographically uniform data. Furthermore, while the technological capability might be present to form server-side aggregations, it does not mean that every data provider will use it. Thus, a more general solution must address the case where a data set is available only by individually accessing each of its files (i.e., URLs) directly.

D. *Extending Metacat to better support aggregations*

At first glance, Metacat and EML provide a robust feature set for addressing the problems presented by the REAP Ocean use-case's searching component. EML can represent aggregations if and when the origin data server cannot provide that capability; Metacat can perform geospatial queries and, for this work, was extended to support temporal search criteria as well.

One limitation with the Metacat/EarthGrid system is the difficulty in processing very large EML files and/or returning very large numbers of EML elements as responses. While the response structure (and EML document structure) are well suited to tens or hundreds of records, with satellite data sets there are often thousands, and sometimes millions, of 'records' returned as part of a single query. This difference, many orders of magnitude in size, placed a strain on the components of the Metacat/EarthGrid system and required that that we build the custom search interface. While it may be impossible to completely address these scalability issues, there may be ways to mitigate them.

Metacat could be extended in two ways that would address these scalability problems and make it a more flexible tool for this kind of search interface. While Metacat queries can be constrained so as to return a subset of the element *type* in a document, it cannot form a subset of individual *instances* of those element types. Thus, Metacat returns *all* of the requested elements in an EML document when *any* of those elements match the search criteria. This means that when the search interface is used to query a limited time range and finds an EML document that contains a single match, it will return all of the URLs for that data set, not just the ones that fell within the query's time range. If Metacat were modified to return only those elements that matched a parametric query (such as those *physical* elements that contain *coverage* elements which fall within a certain time range) then the search interface could eliminate much of the processing of the returned set of URLs.

A second improvement to Metacat would be to preserve the EML element hierarchy in the response it returns. In the current implementation, Metacat 'flattens' the responses making it difficult for the search interface (the recipient of the response) to detect errors that result from missing data in the original EML document. We found that errors did appear in a small fraction of the automatically generated metadata. While it would be best to detect and correct those errors at the source, increasing the overall robustness of the search system would also help trap the cases that will inevitably slip by.

IV. CONCLUSION

We found that building a search interface for the REAP Ocean use-case that used Metacat/EML/Kepler software worked well. Although these SST data are ecological data in the strictest sense, satellite data sets possess different characteristics, and in larger scales, than the ecological data for which Metacat was originally targeted. In spite of these differences, we were able to build a query system using these tools that was not significantly different from other similar interfaces built in the past [9].

While EML imposes no theoretical limits of the number of discrete data objects contained in aggregate data sets, we encountered practical obstacles when using Metacat to query satellite data aggregations containing on the order of 10^4 data objects. The inability to query and directly retrieve specific data object records from within the containing aggregation was the most problematic limitation. Providing support for selective query behavior in a future release of Metacat would eliminate the need for post-processing Metacat search results and would better server the REAP Ocean use-case.

In addition, we found features, such as server-side aggregations, that would have simplified our task were present but underutilized in the software that serves these data. We will investigate ways to simplify the deployment of these data-server-based aggregation techniques and encourage their increased adoption by data providers.

ACKNOWLEDGEMENT

This work was sponsored by NSF grant #0619060. We also thank Dr. Peter Cornillon for help with the Ocean use-case requirements, data server crawling technique and information about satellite data set growth.

REFERENCES

[1] National Research Council of the National Academies, "Environmental Data Management at NOAA: Archiving, Stewardship, and Access," Committee on Archiving and Accessing Environmental and Geospatial Data at NOAA, National Research Council, ISBN: 0-309-11210-9, 2007, pp18-19. http://www.nap.edu/catalog/12017.html.

[2] I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludäscher, and S. Mock, Kepler: "An Extensible System for Design and Execution of Scientific Workflows," 16th International Conference on Scientific and Statistical Database Management, 2004.

[3] "Ecological Metadata Language (EML) Specification," Version 2.1.0, http://knb.ecoinformatics.org/software/eml/eml-2.1.0/index.html.

[4] D. Pennington and W. Michener, "The EcoGrid and the Kepler Workflow System: A New Platform for Conducting Ecological Analyses," ESA Bulletin (Emerging Technologies), 86:169–176, 2005.

[5] D. Barseghian, I. Altintas, M.B. Jones, D. Crawl, N. Potter, J. Gallagher, P. Cornillon, M. Schildhauer, E.T. Borer, E.W. Seabloom, and P.R. Hosseini, "Workflows and extensions to the Kepler scientific workflow system to support environmental sensor data access and analysis", presented at Ecological Informatics, 2010, pp.42-50.

[6] P. Cornillon, D. Crawl, and I. Altintas, "A user based approach to comparing SST fields," Presented at the GHRSST IX Science Team Meeting, Perros Guirec, France, 2008. https://www.ghrsst.org/documents/q/category/ghrsst-science-team-meetings/ghrsst-ix-science-team-meeting/.

[7] F. Monaldo, "Primer on the Estimation of Sea Surface Temperature Using TeraScan Processing of NOAA AVHRR Satellite Data, Version 2.0, S1R–96M–03," The Johns Hopkins University Applied Physics Laboratory, October 22, 1997, p11.

[8] J. Gallagher, N. Potter, T. Sgouros, S. Hankin, and G. Flierl, "The Data Access Protocol – DAP 2.0," NASA ESE-RFC-004.1.1, 2007, http://www.esdswg.org/spg/rfc/ese-rfc-004.

[9] P. Cornillon, J. Chamberlain, and D. Holloway, "The system integrator in distributed data systems: The OPeNDAP Data Connector (ODC)," *in* Sinha, A.K., ed., Geoinformatics: Data to Knowledge: Geological Society of America Special Paper 397, 2006, pp.230-1. doi: 10.1130/2006.2397(18).

CI-Server Framework:

Cyber-Infrastructure Over the Semantic Web

Aída Gándara, Leonardo Salayandía, Aline Jaimes
Cyber-ShARE, www.cybershare.utep.edu
agandara1@miners.utep.edu, leonardo@utep.edu, ajaimes@miners.utep.edu

***Abstract*—Building cyber-infrastructure involves the management of several facets within scientific research, e.g., data, people and processes. Current implementations for sharing research over cyber-infrastructure involve disjoint tools where data centers and portals that do not necessarily focus on an individual research effort are used to house data and metadata. In most cases, this information is only manually searchable leaving little room for automation or emergent knowledge. Since there is little relationship between a data center and an individual research effort, this leaves piecing together the value of a research effort to publications, access to scientists or searching data management centers and data sharing portals. The CI-Server Framework is focused on cyber-infrastructure that supports the documentation of individual research efforts where scientists use tools to seamlessly publish, annotate and comment on their research related resources and where all research information is web-accessible over the Semantic Web. The CI-Server Framework is being used by environmental and geological scientists at the Cyber-ShARE Research Center to describe and document their research.**

Keywords—Cyber-Infrastructure, Semantic Web, Scientific Data Management, Scientific Research Collaboration

I. INTRODUCTION

Nowadays, many scientific research efforts are collaborative, where multiple scientists, often from different domains and even distinct geographic locations, work together toward common research goals. There are many tools that support the sharing of scientific knowledge; scientists use email and chat tools to discuss research amongst two or more collaborators, data management centers are used to publish and share data, and social networking sites are used to discuss, rate and tag shared information. These tools, although they enable sharing of knowledge and information, provide disjointed sharing techniques that are not focused on supporting a research team during or after a collaborative research effort. For example, a research team may use a journal publication to describe their research and outcomes, a data management center to publish the related data and email or electronic meeting notes to capture discussions. Using these tools may provide immediate support to share information but the data they capture is stored in separate locations and may not be accessible to all team members.

Cyber-infrastructure, described and discussed in [1], focuses on enabling scientific teams to work together despite their geographic location yet still supporting the practices of a research organization. The CI-Server Framework is a cyber-infrastructure technology that can be used by scientists to document collaborative research because it enables the sharing of data, metadata, social annotations and discussions about research over the Web. The framework collects related research information as projects, providing a unit of knowledge specific to the research effort. Moreover, the framework emphasizes embedding the CI-Server technology in tools used by scientists as an effort to avoid scientists having to learn the idiosyncrasies of different data management centers and portals. Furthermore, in an effort to enable automated use of the information captured, the framework shares project information as RDF [2], a data model used to describe "things" over the Semantic Web [3].

The CI-Server Framework is currently used by research efforts supported by the Cyber-ShARE Center of Excellence [4], an NSF funded research effort focused on enabling scientific collaboration through cyber-infrastructure. This paper introduces the CI-Server Framework and its support for collaborative scientific research. Section 2 of this paper describes an environmental case study based on a currently active Cyber-ShARE research effort. Section 3 discusses details of the CI-Server Framework while section 4 highlights how the content collected in a project can enable automation. Section 5 discusses related and future work and Section 6 discusses some conclusions.

II. ENVIRONMENTAL CASE STUDY

Eddy covariance methods [5] are being used by the Systems Ecology Lab (SEL) at the University of Texas at El Paso to study land-atmosphere interactions in a desert ecosystem to better understand the process of desertification that is affecting rangelands worldwide. The station is located at Jornada Basin Experimental Range in Las Cruces, New Mexico. Investigators at SEL calibrate, operate, and maintain the instrumentation on a flux tower. They also retrieve, process, and archive the data using customized methods and infrastructure which have been developed after combining scarce information from literature, fellow researchers, manufacture's procedures, and National and International networks guidelines.

Eddy covariance methods are some of the most direct methods to measure the vertical turbulence that drives the

mass exchange of heat, water vapor, and carbon within the atmospheric boundary layer) [6, 7, 8], however, deploying eddy covariance towers is time consuming and costly, data processing is mathematically complex and the learning curve is high. Hence, eddy covariance tower deployments typically are planned to be used for long-term studies. In order to justify such investments, investigators are usually motivated not only to answer specific scientific questions about the region where the flux tower is deployed, but also to share the data with the broader community, e.g., through the FLUXNET community [9].

In the case of the eddy covariance tower at Jornada, investigators at SEL are maintaining the data in the raw formats offered by the instrumentation used in the field. In addition, they use specific software packages that are available from the eddy covariance community to preprocess flux calculations from the raw data. Due to the nature of eddy covariance methods, it is inevitable that failures in the environmentally exposed infrastructure will yield gaps in the datasets being measured; furthermore, having a complete dataset is crucial to capture the fast changing environmental conditions of the region in a day cycle. As a result, a critical part of the processing of eddy covariance data is the gap-filling process, by which specialized algorithms are fine tuned according to the specific environmental conditions of the particular eddy covariance site to identify gaps and fill them with meaningful values. The gap-filled data, called the corrected data, is also archived, along with flux calculations derived from both the raw data and the corrected data.

Data from the Jornada eddy covariance site is retrieved in near real-time using a WIFI connection established through the Jornada headquarters and transmitted over the Internet to SEL file servers. As a backup, researchers swap internal logger storage cards, and a laptop with an external hard drive is used to extract the data from the on-site data loggers and physically transported to SEL file server. At the time of storing data in the file server, a de-duplication routine has to be performed manually to reconcile the data received through WIFI connection and that received by the external hard drive dump. Using conventions conceived by SEL investigators to store the data, the SEL file server is organized mainly by date, and additional spreadsheet documents and readme files in text format are created to capture additional ancillary data.

From the perspective of eddy covariance dataset users, accessing datasets from the Jornada eddy covariance tower requires contacting investigators at SEL. Although the data can be made accessible directly over the Web, personal contact is still necessary to understand the idiosyncrasies specific to SEL to store the data and ancillary data, as well as to describe the specific gap filling routines used. This supporting information is often maintained in a combination of notebooks, emails and spreadsheets that are kept separately and in addition to the data at the SEL file server.

From the point of view of colleagues operating other eddy covariance sites, personal interaction is needed to share calibration records, field data entry forms, and other information required to implement, maintain, and improve site operation. Furthermore, these personal interactions usually result in unstructured artifacts that even if accessible by the community, are difficult to find and contribute to. As a result, SEL scientists are using a combination of techniques to document and share their research with other scientists.

III. CI-SERVER FRAMEWORK

Often times, when sharing data, scientists will choose to place their data on some type of externally available Web server, e.g., a portal or data management center. The metadata is chosen by the site owners and the data itself is not normally formatted for viewing, e.g., it is an XML file or a binary file, thus scientists are limited in how they describe a dataset's relationship to a research effort. That is, how data is related to a research effort or an organization is second to the rules by which they publish their data.

The CI-Server Framework was created with the goal of understanding how to support scientists in documenting and sharing ideas and knowledge about collaborative scientific research. The SEL scientists must describe their research for the purposes of discussion, publication and overall understanding. As noted in the case study, the data itself is stored at the SEL file server and the documentation of process is maintained separately. It was important when identifying the characteristics of the CI-Server Framework, to avoid changing the scientific research practices; rather the focus is on enhancing the practices to work within the Cyber-ShARE cyber-infrastructure. Thus, the CI-Server Framework approach is to help scientists describe their research effort electronically without having them focus too much on how they will share their data. Ultimately, the goal is to provide scientists the ability to annotate and share specific details about a research effort so as to enable understanding, automation and reuse.

The framework consists of a Drupal–based [10] Web server that supports the collection and management of information, a Java-based client API that exposes server functionality and various tools and applications that make use of the server data, in particular to publish, retrieve and discuss data collected on the server.

A. The CI-Server

The CI-Server, the Web server in the CI-Server Framework, is built from a Drupal 6 Content Management Server install and additional contributed modules from the Drupal Community [11], e.g., Taxonomy, CCK, Services. Modules in Drupal are PHP extensions that provide additional functionality in a Drupal installation. The CI-Server is implemented in a server module, that controls functionality like menus, views and projects on the server, and a services module that provides the server side functionality of the API. The API is implemented as XMLRPC services. The CI-Server considers two types of information to manage, a content type, these are the main Drupal resources, and attachments to content types as either files or links. Identifying content types and a related attachment supports the fact that although the actual files or links referenced by a scientific research effort are unique, there are consistent attributes that can be shown in Web forms on the CI-Server for a specific Drupal content type. Attachments are defined by adding fields to content

types. A file attachment is a file that is physically loaded to the CI-Server. A link attachment is a reference to some resource that is located elsewhere and accessed via a URL. Since a Drupal server can have many content types, an administrator can configure which content types are accessible to upload and download thru the CI-Server services module. All content types and files located on the CI-Server are accessible through an assigned URL.

Projects are used to group related content for a research effort. Projects are implemented as a Drupal content type and a taxonomy tag. In Drupal, a comments are related to content types, thus the project content type is used to collect project level comments for a research effort. Using a taxonomy to tag resources as related to a research effort provides flexibility; allowing for the same resource to be referenced by more than one research effort. We should note that in the initial CI-Server implementation, projects were organized by placing related resources in a single directory. We learned that this design severely limited how administrators could organize their data and how scientists could reference their data. Now, data is categorized by the node type defined on the server, accessed via the URL and grouped with project tags.

We learned, from working with scientists [12], that the reasons why scientists choose certain characteristics in their research steps is just as important as the results. Moreover, why research was conducted helps to understand the overall research effort, something that could affect the long term reuse of the research results. CI-Server leverages the comments that can be added to Drupal content types to capture scientist's comments about scientific research. Project comments can be accessed from content type views as well as the client API enabling a scientist-centered description of scientific research.

Figure 1 shows the resource view for the EddyCovariance Project, described in the case study of Section 2. This view provides a list of all content, resources, that are included in this project, organized by content type, e.g., PMLJs, SAWs, UDATAs and WDOs. From this view, a user can open the file or link of a content type in a relevant tool or the content type can be opened to view the Drupal fields for the content type. The tool to open or visualize a project resource is configured by a CI-Server administrator. The SEL scientists are currently using a workflow tool to describe the process by which they are conducting their research, these are reflected by the SAW and WDO resources, they have published data that has been created from conducting research and they have published some provenance files, i.e. PMLJs, that collect knowledge about how specific data was created.

As opposed to scientists finding a Web server to place their data, being limited to the type of data that can be uploaded, conforming to the organizational rules of the Web server or searching for relevant data all over the Web, the project data on the CI-Server is accessible as a unit and the content views can be further configured in Drupal for all CI-Server users or for specific projects.

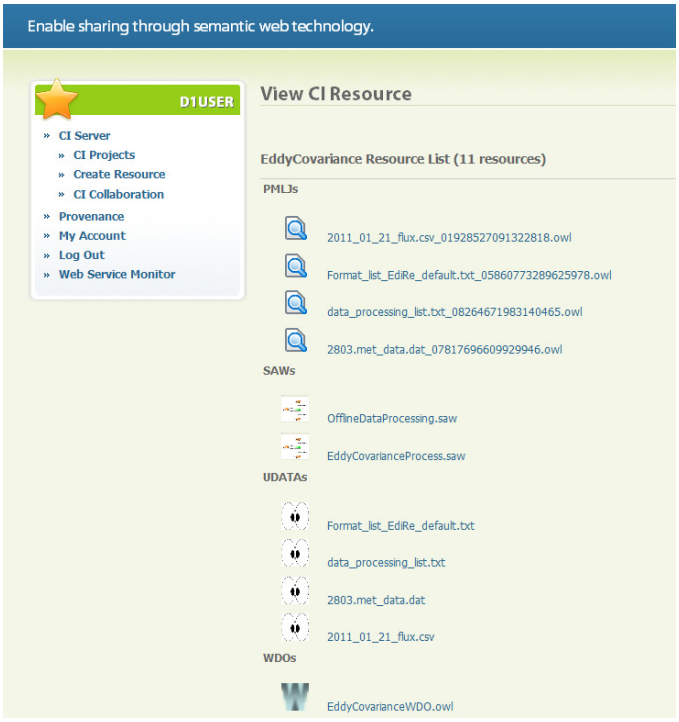


Figure 1. The CI-Server Framework supports scientific teams in describing their research by managing related resources in projects. This figure shows the resource view, a list of resources, of the EddyCovariance project described in the case study in Section 2.

B. The CI-Client API

When scientists consider the management of their scientific data, there is often a concern as to the investment that incorporating new tools and technologies will take and, more importantly, the potential delay or alterations to their research process. One goal for the CI-Server Framework was to avoid the issues scientists have with conforming and learning different rules for publishing and accessing their data. Ultimately these rules are important to support collaboration, but they are details that have little relevance during a research process.

The CI-Client API is an API built in Java that provides an interface to the services defined in the CI-Server services module. Through instrumentation of this API in existing tools that our scientists are using, scientists share and collaborate within the tools that they know. There are four classifications of calls. Some perform administrative type actions like creating projects, finding resources, logins to the system or obtaining user specific information. There are calls that publish and upload information to the server and calls that download information from the server. Finally, there are user interface calls that facilitate server functionality, for example, there is a view that lists known servers, a view for logging on to a server, a view for selecting a project on a server.

Since the CI-Server Framework is a Web-based framework, it must be sensitive to the fact that all resources are uniquely addressable. When a user places files on a local system, files have unique names. The same occurs on a Web server except that the CI-Client API must obtain specific URL

names from the CI-Server in order to return appropriate URL names. This alleviates two issues. First, by obtaining these rules from the CI-Server, the CI-Client can create resources on the CI-Server without violating its naming, second, if a client tool creates documents that are internally linked, the correct URLs can be used to create these links avoiding some type of translation later, when the file is published on a server.

As a result of the CI-Client API, research efforts supported by the Cyber-ShARE Center have been able to share more data over the Web, with little concern for the logistics of uploading and downloading at one particular Web server. They have published this data from tools that have been instrumented with the API, thus they focus on describing and conducting scientific research, not learning a Web server interface.

C. *Making Use the CI-Server Framework*

Currently, there are two CI-Server implementations used by the Cyber-ShARE Research Center, one is a production system and another is a test system used to explore server enhancements. There are two more implementations planned and there have been additional installations serving to demonstrate the CI-Server Framework functionality at different user sites. The CI-Client API has the capability to connect to each of these servers and access the data published at a server; client tools that are instrumented with the CI-Client API can establish a connection to any CI-Server. There are currently 6 active research efforts that are publishing and accessing CI-Server content; spanning multiple domains. For example there is research on geological crustal modeling, environmental eddy covariance, health related issues as well as support for more general content like provenance traces and publications. Scientists are publishing this data using CI-Server clients and accessing data either through the URLs or the CI-Client API interface. There is no restriction to the types of documents published on a CI-Server. Currently there is a collection of publications, data sets, OWL files, xml documents, and more.

There are also a variety of Cyber-ShARE tools [1] that integrate with the CI-Server Framework. The CI-Desktop is a Java tool that is provided with the CI-Server Framework. Using the CI-Desktop interface, a user can connect to any CI-Server, browse its contents and upload or download resources. Other Cyber-ShARE tools are not provided with the CI-Server Framework but have been enhanced to make use of the CI-Server Framework, e.g., DerivA has been instrumented with the CI-Client API to capture and publish provenance behind manual scientific processes, ProbeIt uses the links within RDF documents to show visual graphs of provenance described in PML[13], a provenance data model. SPARQL-PML is a triple-store query engine that crawls, loads and reasons over all RDF content that has been published at a single CI-Server. WDOIt! is a tool used to describe scientific processes, this tool creates process specifications, also called workflows, encoded in OWL[14]. In some cases, these tools already existed but they were lacking the capability to share in a Web-based environment, in other cases, these tools were created taking advantage of or contributing to the information shared on a CI-Server.

[1] http://trust.cs.utep.edu

Due to the complexities in how SEL scientists are conducting their research, these scientists are documenting their research using tools that have been instrumented to interact with the CI-Server Framework. WDOIT! is a tool that provides a graphical user interface to document processes that may be automated or human-driven. SEL scientists are using WDOI! to capture a scientist's understanding of a process, i.e., focusing on what, when, and why activities are performed to achieve a scientific outcome, while disregarding technical nuances, like executable knowledge of how activities are performed. Before the CI-Server Framework, WDOIt! was another disjoint tool that researchers used to document scientific processes, resulting in silos of process specifications that would be shared at a later point, e.g., published at a Web server. Initially, scientists would build process specifications on a local computer system; where the process specifications would reference resources on the local file system. From a local system, there was no mechanism to share the documents aside from copying them to a shared location or sharing them via email. Migrating WDOIt! process specifications to other locations for sharing is a tedious task because internal references, using the WDOIt! OWL encodings, to other resources would usually break. This would require a manual update to fit the references to resource locations on the new system. Using email to share and discuss process specifications would also run into issues. For example, losing reference to which attached version was the master or only including some team members in the discussion. By instrumenting WDOIt! with the CI-Client API, scientists can immediately publish process encodings for collaboration on a CI-Server. The internal links of the processes are resolved by the CI-Server, the process files are related to a project and all content is made web-accessible, i.e., assigned a URL. WDOIt! also uses a CI-Client API interface for submitting comments about a process and these comments are published at the server, annotating the resource using Drupal comments. Other users can use the CI-Server to view all resources related to the project, via the CI-Server resource view (see Figure 1) and they can see the graphical representation of the process by selecting the view link for a resource. As a result, sharing information is a byproduct of using tools like WDOIt!, that make use of the CI-Server Framework; not a subsequent step that scientist's are responsible for when they need to share their research.

IV. ENABLING AUTOMATION

There are several examples on why structured semantic data like RDF and OWL can be useful to sharing information, e.g., information integration, ontology alignment and tagging [15]. Using a RDF- based structure in data can also help with searches, because of consistent terminology and categorizing data. More importantly, for the CI-Server Framework, structured data enables machines, i.e. software agents, to understand data, a quality of content published over the Semantic Web.

Several tools already exist that take advantage of RDF data. For example, RDF browsers are tools that load and provide browsing views for any RDF dataset without having any prior knowledge of the content, aside from its description in RDF. Moreover, they allow users to follow links to other

related data, preferably also described in RDF. This capability is a result of the Linked Open Data [16] effort that is focused on making links between semantically described Web content then allowing machines to resolve those links; relieving users from having to find them or Web pages from having to hardcode them. Tools that have aggregators, for example Sindice [17], can load RDF data from multiple RDF data locations. SPARQL [18], an RDF query language, can then be applied to the entire RDF dataset because, despite their actual physical location, all the data is in the same structure, i.e., an RDF triplestore. In this way, there is access to more related data and reasoning is enabled over more knowledge.

The information that is collected in a project within a CI-Server is a nucleus of information available about a research effort. In order to make this information useful outside the research effort, the CI-Server provides an RDF view for project resources. Through a URL, software agents can access RDF descriptions of a project, its comments and all project resources. RDF data about a CI-Server project can be loaded into an RDF browser and users can see how this data links to information not necessarily included in the project. Similarly, one or more RDF project datasets from a CI-Server can be loaded into an aggregator and SPARQL queries can be created to answer queries about the project.

The potential here is that although we do plan on exposing additional knowledge from the RDF that is generated for a project, see the next section for Future Work, gaining additional knowledge about a research effort is not restricted to the functionality of a CI-Server. The CI-Server, in servicing multiple research efforts, is not equipped to understand all data. This open and structured model of the CI-Server Framework to provide URLs for resources, group research efforts by project and expose projects as RDF, enables further automation by external software agents.

V. Related and Future Work

Portals and data management centers have been mentioned throughout this paper because they are currently used by scientists to publish and share their data. The documentation of a research process as graphical scientific workflows is, in our view, an effective way to convey process for understanding [19]. Two related implementations use executable scientific workflow systems to conduct scientific experimentation and support collaborations through social Web portals, i.e., they allow workflows to be published on Web portals for further sharing and discussion of scientific processes. myExperiment is a web portal used to publish, discuss and rate scientific workflows [20]. Different types of workflows can be published as a single workflow package giving users of the system access to reusable workflows and related data. Users can download packages, execute them on their local system and push changes back onto the server. Users can also discuss their opinion of workflows and rate workflows. Another scientific workflow based portal is CrowdLabs, a social repository for VisTrails workflows [21]. In this repository, users are able to access, discuss and rate workflows. VisTrails workflows can be opened and modified locally, i.e., on the client system using a local VisTrails application, and changes are pushed back to the server. CrowdLabs manages projects where scientists can discuss specific VisTrails workflows with other scientists.

As opposed to the CI-Server Framework, these two implementations support workflows and workflow discussions, not necessarily research efforts. Both have a predefined set of file types that can be published. The data, e.g., myExperiment packages and VisTrails workflows, and associated user comments, are not openly available RDF resources. myExperiment provides a server based SPARQL endpoint, where SPARQL queries can be executed to retrieve RDF. By using a predefined ontology, the myExperiment portal can control how data is described and therefore provide some level of search. Crowdlabs requires that workflows be accessed manually as projects and there is little openness to data unless the user is an authenticated user in the system. Neither portal seems to provide the ability to integrate client tools to publish data or comments or to retrieve information, e.g. there is no API that integrates client tools with the functionality available at the portal. As a result, scientists must understand each portal specifically and manually, including menus and data organization, if they want to interact with the portals. Finally, these two portals seem to be single implementations whereas the CI-Server Framework is meant to be replicated; site administrators can download the CI-Server components and setup a more specific Drupal-based server for their needs. Cyber-ShARE client tools can be downloaded from the Cyber-ShARE website or technologists can instrument scientist specific tools by downloading and integrating the CI-Client API.

Although the CI-Server will unlikely be able to support reasoning for all research efforts individually, we believe that there is some insight we can provide for the value of a focused project-based RDF dataset. Our goals are to provide views based on queries into the RDF data that is generated for a project as well as browsing views, where users can follow links to other relevant data in the Linked Open Data cloud. We believe that with these views, scientific teams and outside parties interested in the research can gain a better understanding of the research effort versus what they would find if they were to perform searches of this data over the Web.

VI. Conclusions

Although there has been no formal user evaluations of the CI-Server Framework, there has been a growing dependence on the framework from Cyber-ShARE Center technologists and scientists. Where before the Center's scientists had silos of information, scientific teams are creating projects and publishing more information describing their research. The framework, since its inception two years ago, now has over 56 projects where various users have found it useful to publish data, formats, metadata, publications, workflows and provenance as Web accessible entities, i.e., having URLs. The main production CI-Server houses over 5000 resources with corresponding attachments, most of which have been published using tools instrumented with the CI-Server API, and provides service to approximately 15 to 20 users who logon to actively use the server. This does not include users

who access the data via URLs. For Cyber-ShARE scientists, this exhibits the benefits to embracing the scientific process that these scientists are engaged in and the effect of seamlessly integrating technology in the tools that scientists use, as opposed to requiring them to find mechanisms for sharing.

The CI-Server Framework has a different focus from existing scientific data management centers and data sharing portals, namely to support the individual research effort. Team support is achieved by allowing scientists to connect to a CI-Server through tools that enable them to describe research without having to understand the details of uploading or downloading data. As a result, scientists minimally alter the process of documenting their research because they can use familiar tools that capture this information electronically. Through grouping related information as projects, this information is maintained as a unit to help describe a single research effort and can therefore be shared as a single project via a resource view on a CI-Server. Making the information on a CI-Server Web-accessible via URLs enables future users to reference project resources individually or the project as a whole.

Furthermore, the CI-Server Framework is enabling automation and reuse by assuring that all resources are available as RDF. As a result, humans can rely on semantically enabled tools to help make use of Web content, in particular when the information is obscure or massive. For the SEL scientists, it is our expectation that as they add additional resources and comments to the EddyCovariance project, leveraging semantically enabled technologies should provide additional understanding of the research effort.

ACKNOWLEDGMENT

This research was partially funded by the National Science Foundation under CREST Grant No. HRD-0734825. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors would like to thank Dr. Paulo Pinheiro da Silva and Dr. Craig Tweedy for their invaluable advice and direction on this research.

REFERENCES

[1] NSF CyberInfrastructure Council, NSF's CyberInfrastructure Vision for 21st Century Discovery, National Science Foundation, March 2007.

[2] Resource Description Framework (RDF) Model and Syntax Specification , Ora Lassila, Ralph R. Swick, Editors. World Wide Web Consortium Recommendation, 1999, http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

[3] Berners-Lee, Tim; James Hendler and Ora Lassila (May 17, 2001). "The Semantic Web" . *Scientific American Magazine*.

[4] Cyber-ShARE Center of Excellence (2011). http://cybershare.utep.edu.

[5] D.D. Baldocchi and T. P. Meyers, On using eco-physiological, micrometeorological and biogeochemical theory to evaluate carbon dioxide, water vapor and gaseous deposition fluxes over vegetation. Agricultural and Forest Meteorology, volume 90, 1998, pp. 1–26.

[6] Baldocchi, D., B. Hicks, and T. Meyers. 1988. Measuring biosphere-atmosphere exchanges of biologically related gases with micrometeorological methods. Ecology 69, pp. 1331-1340

[7] Lee, X., W. Massman, and B. Law. 2004. Handbook of Micrometeorology. Kluwer Academic Publishers, The Netherlands, pp. 250

[8] Burba, G.G., and D.J. Anderson, 2010. A Brief Practical Guide to Eddy Covariance Flux Measurements: Principles and Workflow Examples for Scientific and Industrial Applications. LI-COR Biosciences, Lincoln, USA, pp. 211

[9] About Fluxnet (2011). http://www.fluxdata.org/SitePages/AboutFLUXNET.aspx

[10] Drupal (2010). http://drupal.org

[11] Drupal Community and Support (2011). http://drupal.org/community

[12] A. Gandara, G. Chin Jr., P. Pinheiro da Silva., S. White, C. Sivaramakrishnan, T. Critchlow. Knowledge Annotations in Scientific Workflows: An Implementation in Kepler, To Appear in Proceeding of Scientific and Statistical Data Management Conference, Portland, Oregon, July 20-22, 2011.

[13] PML, Proof Markup Language (2010). http://tw.rpi.edu/portal/Proof_Markup_Language

[14] Deborah L. McGuinness and Frank van Harmelen. OWL Web Ontology Language Overview. World Wide Web Consortium (W3C) Recommendation. February 10, 2004. At http://www.w3.org/TR/owl-features/.

[15] N. Shadbolt, W. Hall, and T. Berners-Lee, "The Semantic Web revisited," IEEE Intelligent Systems, pp. 96–101, May/June 2006.

[16] Christian Bizer, Tom Heath, Tim Berners-Lee: Linked Data - The Story So Far. In: International Journal On Semantic Web and Information Systems, Vol. 5, Issue 3, Pages 1-22, 2009.

[17] Sindice The Semantic Web Index (2011) http://sindice.com

[18] Prud'hommeaux, E., Seaborne, A.: SPARQL Query Language for RDF. W3C Working Draft, 4 October 2006. Available at: http://www.w3.org/TR/2006/WD-rdf-sparql-query-20061004/

[19] A. Gates, P. Pinheiro da Silva, L. Salayandia, O. Ochoa, A. Gandara, N. Del Rio, Use of Abstraction to Support Geoscientists' Understanding and Production of Scientific Artifacts. (2008) Editors: G. Randy Keller, Collection: Cambridge University Press Book: Geoinformatics: Cyberinfrastructure for the Solid Earth Sciences. Bibliography: School of Geology and Geophysics, University of Oklahoma, Chaitanya Baru, San Diego Supercomputer Center, University of California, San Diego.

[20] D. D. Roure, C. Goble, and R. Stevens. The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows. Future Generation Computer Systems, 25:561–567, 2009.

[21] P. Mates, E. Santos, J. Freire, C. T. Silva. CrowdLabs: Social Analysis and Visualization for the Sciences, To Appear in Proceeding of Scientific and Statistical Data Management Conference, Portland, Oregon, July 20-22, 2011.

Geo-Seas: a pan-European marine geoscientific e-infrastructure

Helen Glaves[1], Dick Schaap[2]
[1] British Geological Survey
[2] MARIS
hmg@bgs.ac.uk; dick@maris.nl

***Abstract*—Despite the fact that there are large volumes of geological and geophysical data available for the marine environment it is currently very difficult for users to locate and access these datasets or use them in an integrated way. This is due to the use of different nomenclatures, formats, scales and co-ordinate systems not only between individual countries, but also within the same country between different organizations. In an attempt to overcome some of these difficulties the Geo-Seas project is developing an e-infrastructure for the delivery and exchange of marine geological and geophysical data. This infrastructure is made up of 26 data centres in 17 European coastal countries and includes research and academic institutes as well as a number of national geological surveys.**

Keywords—metadata; e-infrastructure; data delivery; marine geoscience

I. INTRODUCTION

The main objective of the Geo-Seas project is to provide direct user-access to harmonised marine geological and geophysical metadata and datasets through the development and use of common standards, vocabularies and methodologies [2, 3]. The project also aims to enhance interoperability with other data types and infrastructures such as those used in the wider earth sciences community [4]. Geo-Seas is also underpinning key European directives such as INSPIRE (Infrastructure for Spatial Information in Europe), which is developing standards and a structure for delivering integrated spatial information services, as well as international initiatives such as the Global Monitoring for Environment and Security (GMES) and the Global Earth Observation System of Systems (GEOSS), both of which are encouraging the provision and exchange of environmental data and information.

Geo-Seas is adopting and adapting the technologies developed by the related SeaDataNet project for use with geological and geophysical data types. SeaDataNet has implemented an e-infrastructure for the management of oceanographic data which is based upon a distributed data model with each individual data centre responsible for the management and delivery of their own data sets. Each data centre also provides metadata records for their locally held data sets to the centrally managed Common Data Index metadatabase. This metadata then provides the link between the Data Discovery and Access Service and the data held by the individual data centres for the purposes of data discovery and download.

The Geo-Seas project is now implementing a similar model for geoscientific data in order to create an e-infrastructure which allows a range of users including researchers, academics and policy makers to directly access harmonised marine geological and geophysical data sets through a single dedicated portal which is available via the project website at http://www.geo-seas.eu.

II. BACKGROUND

The project is building upon the work of the SeaDataNet project which has created an e-infrastructure for the delivery of oceanographic data throughout Europe. Geo-Seas is now adopting and adapting the architecture, methodologies and technologies developed by SeaDataNet for use with geological and geophysical data. This has resulted in the development of a joint e-infrastructure covering both oceanographic and marine geological and geophysical data which in turn has facilitated the development of multidisciplinary science through the creation of interoperable data sets for use in both ocean science and the wider user communities. Geo-Seas has also incorporated the work done by a number of earlier European Commission-funded projects including EUSeaSed and SEISCAN, both of which created extensive marine geoscience metadatabases. These pre-existing metadata catalogues have been used as the basis for the development of the Geo-Seas metadata standards and they have also been upgraded to conform to the ISO19115 for incorporation into the Geo-Seas metadatabase.

The re-use of the SeaDataNet methodologies and technologies, including both the architecture and middleware components where appropriate, to interconnect the geological and geophysical data centres, will enable the integration of geological and geophysical datasets with other oceanographic data currently managed by the SeaDataNet data centres. Not only will this avoid unnecessary duplication of effort within the two projects but will also allow the development of a common approach to marine data management across Europe which can potentially be extended to the wider international community [4].

III. METHODOLOGY

A. Metadata

The individual data centres are required to create metadata for their datasets with each one having a metadata record which references the data at the file level. This metadata conforms to the ISO 19115 standard but, in order to include the additional information required for oceanographic data the SeaDataNet project created an enhanced metadata schema, the Common Data Index (CDI). This schema has been further extended and adapted to accommodate the specific requirements for the delivery of geophysical data, and in particular seismic data using the Observations and Measurements (O&M) and SensorML schema. The CDI schema has also been upgraded to include detailed tracks and polygons for referencing geophysical data. This has been achieved using the Open Geoscience Consortium (OGC) compliant Geography Mark-up Language (GML) which includes an option for additional service bindings to provide a linkage to the viewing services which are also required for this project.

It has been shown in previous projects that the use of common vocabularies is essential to ensure consistency and interoperability [1] [5]. For this reason a set of common vocabularies is being used for the population of the keyword fields as part of the creation of the standardised metadata. The common vocabularies, originally established for use in the SeaDataNet project, are widely used throughout the oceanography community and they are now being updated by the Geo-Seas partners to accommodate the specific requirements of geoscientific data.

The update of the common vocabularies has been undertaken in two phases. During the initial phase a domain expert group was established which included representatives from both the project and other relevant initiatives. This group was tasked with evaluating and extending the pre-existing vocabularies to include the geoscientific terms required to populate the discovery metadata keyword fields.

During the subsequent data population phase of the project further extension of these common vocabularies has been necessary to include additional terms identified by the partners engaged in these population activities. As a result, the updating of these common vocabularies is currently an ongoing task with new terms being added to the relevant vocabularies as necessary.
In order to maintain the integrity of these common vocabularies all requests for additional terms to be added to these lists must be validated and approved. This is achieved through an international vocabulary content governance group (SeaVoX)
https://www.bodc.ac.uk/data/codes_and_formats/seavox/

In order for the partners to be able to create the standardised Common Data Index (CDI) records each data centre has been provided with a Java® based software tool, MIKADO, which was originally developed by the SeaDataNet project to facilitate the creation of the required metadata. This tool has also been adapted and updated for use with geosciences data. The MIKADO tool is used to generate the CDI metadata records, which are in an XML format, directly from local partner databases either automatically as a batch job or manually. Each individual partner must have first carried out a local mapping exercise between their local database and the Common Data Index schema and also the common vocabularies. The MIKADO tool references these vocabularies using local web services to get up-to-date lists as part of the metadata generation process. Once the partner has created the CDI metadata this is then loaded to a centralised project database which is utilised by the Geo-Seas portal for the data discovery services.

B. Data delivery and exchange

To facilitate the delivery of the data in a standardised format the project partners have developed an agreed set of data delivery and exchange formats which have been adopted by all of the data centres (Table 1). These agreed formats have been chosen as they are the most commonly used standards within the oceanographic community and are also either already used or can easily be adapted for geosciences data. They include Ocean Data View (ODV) which is an ASCII format widely used in the oceanographic community for profile, time series and trajectory data; modified NetCDF (CF) which is a data exchange format commonly used for gridded data sets and SEG-Y which is generally used for the delivery of geophysical data

Each partner must make their data files available on a local server in the format agreed for each data type and create a link between them and the associated CDI metadata record using the software tools provided. In order for these data files to be accessible via the Geo-Seas portal each data centre is also required to install the Download Manager software tool. The Download Manager can be used in one of two modes. The

TABLE I. EXTRACT FROM LIST OF AGREED DATA DELIVERY AND EXCHANGE FORMATS

Data Type	**Delivery Format**
Geological data (point)	ODV & GeoSciML
Geological data (gridded)	NetCDF
Gravimetry (tracking)	ODV
Gravimetry (gridded)	NetCDF
Bathymetry (tracking)	ODV
Bathymetry (gridded & swath)	NetCDF
Borehole	ODV & GeoSciML
Heat Flow	ODV
Magnetic (tracking)	ODV
Magnetic (gridded)	NetCDF
Seismics (digital data)	SEG-Y
Seismics (scanned images)	TIFF / PNG
Seismics (navigation)	UKOOA
Side scan sonar	XTF

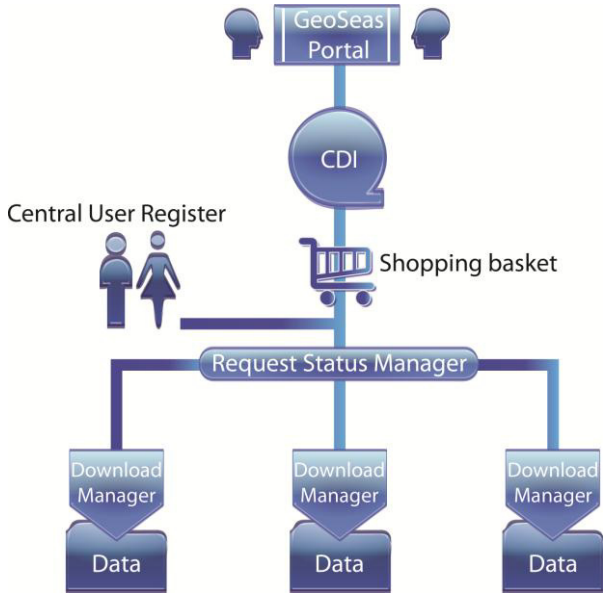


Figure 1. Geo-Seas data portal access

data centre can convert the relevant data files to the agreed formats and store these on a local server which can then be accessed by the Download Manager. Alternatively the data centre can store the data in their local database in these delivery formats which can then be downloaded using direct database calls via the Download Manager.

Once the data centre has created the CDI metadata records and loaded them to the central metadatabase, converted the associated data files into the agreed formats and installed all of the necessary software components, they can then be connected to the Geo-Seas e-infrastructure as a fully functioning data centre. Some of the data centres will also install an ancillary set of software tools specifically for the delivery of the geophysical data. However, these are not essential for an organisation to be a fully functioning data centre and are only required for specific data types which are not held by all data centres.

Once a data centre is connected to the e-infrastructure, the end user can then access and download the data holdings of that data centre. The Geo-Seas portal (Fig. 1) allows users to search through the metadata catalogue, find the data they need (Fig. 2), assess its suitability for their particular purpose and then either download the data directly from the data centre or place an order for the data according to the access and use restrictions which have been put in place by the data supplier. The datasets that can be accessed and directly downloaded by an individual user will be determined by the status of the registered user.

The Geo-Seas portal and discovery metadata services are public domain but in order to use the data download service an individual must first become a registered user. As part of the registration process the user must agree to abide by the conditions of the data user licence and is assigned a status according to their affiliation (academic, commercial, etc). Each time the user places a request to download a data set via the Geo-Seas portal they are required to log-in. When ordering any data the status of the user will be verified and, where there are access and use conditions, the user may be referred directly to the data centre to negotiate the terms and conditions for the use of a data set before delivery can proceed.

Once the end user has placed a request for data from a data centre the progress of that order can be monitored by the user via the Geo-Seas portal. The Request Status Manager (RSM) application (Fig. 1) controls the ordering and delivery of the data from the separate data centres as no data is held centrally. All of the raw data remains under the management of the data centres that hold the data. This allows the individual data centres to retain responsibility for managing their data holdings locally whilst optimising the delivery of these data sets to the wider user community.

IV. SUMMARY

The Geo-Seas project is currently in the installation and population phase with each of the project partners having installed the software tools necessary in order to become a fully operational data centre and part of the e-infrastructure. They are also engaged in the creation of the metadata for their respective data holdings and making the associated data sets available on local servers in the agreed delivery formats in order for them to be accessed via the centralized Geo-Seas portal. Each of the data centres is now uploading metadata to the Common Data Index metadatabase and of these data centres more than half are now fully connected to the portal with in excess of 41 000 data sets having already been made available via the Geo-Seas portal. Over the coming months this figure will increase as the remaining data centres come on-line and those data centres that are already connected making additional datasets available. The data sets currently available cover a wide range of different types including geological data derived from both observation and analysis of geological samples e.g. grain size,

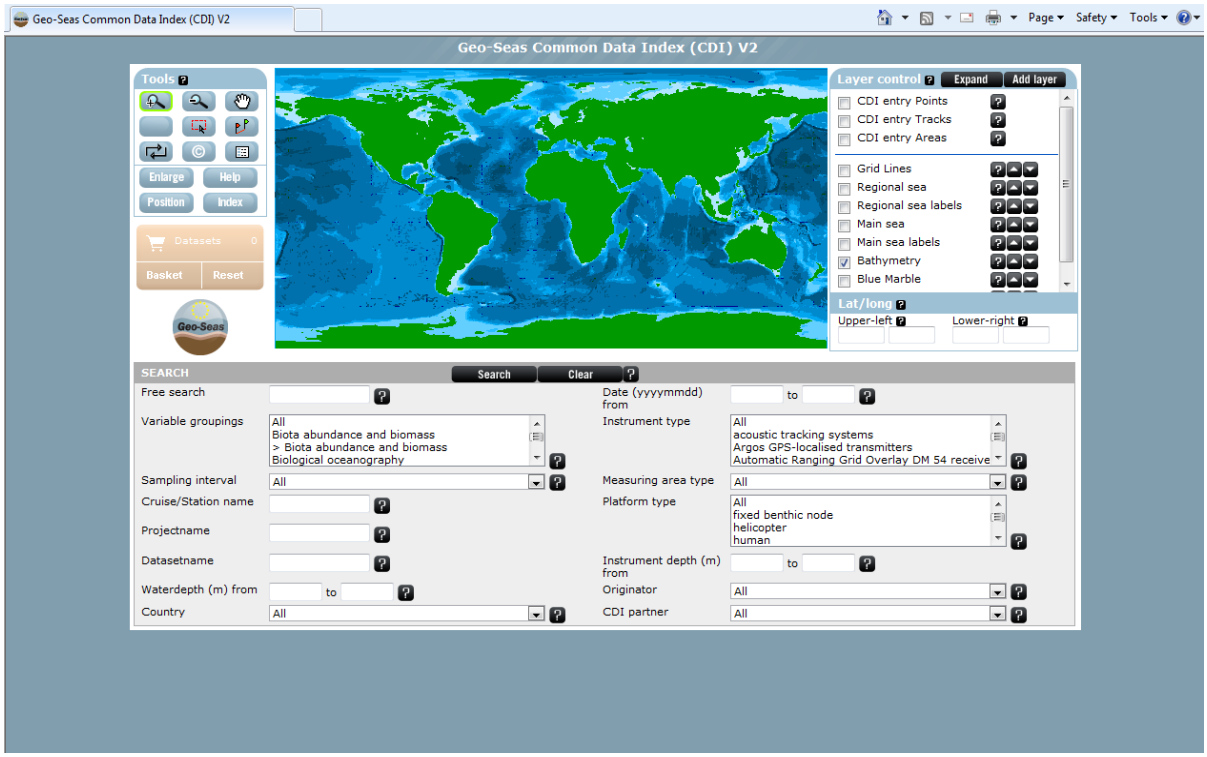


Figure 2. Geo-Seas Common Data Index metadata discovery service

geochemistry; as well as bathymetric data. Additional data types including geophysical data such as a multibeam and side-scan sonar will also be made available over the coming months.

The development and implementation of the Geo-Seas portal has provided users with a single point of access for the discovery and delivery of harmonised marine geological and geophysical data within Europe. This is leading to a significant improvement in the locating, accessing and delivery of a range of interoperable marine geoscientific data sets. Users can find the data they need and assess its suitability for a particular purpose and then either download the data directly or place an order for the data dependant on the status of the registered user (academic, commercial etc.) and also the volume of data requested.

The implementation of common standards and methodologies is also leading to improved interoperability of marine geological and geophysical data with other data types and data products from other disciplines, organisations and between countries. It is also allowing the development of multidisciplinary marine science within Europe on a whole-basin scale.

V. REFERENCES

[1] L. Bermudez , Graybeal, J., Isenor, A., R. Lowry, R. and D. Wright, D. "Construction of marine vocabularies in the Marine Metadata Interoperability Project," In: Proceedings of the Marine Technology Society / Institute of Electrical and Electronics Engineers Oceans Conference, Washington, D.C, 2005

[2] H. Glaves and Graham, C. "Geo-Seas - a pan-European infrastructure for the management of marine geological and geophysical data," EGU General Assembly 2010

[3] C. Graham and Schaap, D. "The GEO-SEAS project: a European network for marine and ocean geological and geophysical data,".*Baltica,* 22 (1). pp10, 2009.

[4] S. Miller, Glaves, H., Carbotte, S., Arko, R., Chandler, C., Clark, D., Sweeney, A. and Stocks, K, "Rolling Deck to *Reposit*ory (R2R): Opportunities for US-EU Collaboration," EGU General Assembly 2011, ESSI2-3939

[5] D. Wright, Watson, S., Graybeal, J. and Bermudez, L. "Making Scientific Data Sets Easier to Find, Access, and Use," Eos Trans. AGU, 86(50), 2005

ACKNOWLEDGMENT

The author would like to acknowledge the 28 consortium partners who are undertaking the development of the Geo-Seas e-infrastructure. This work is being co-funded by the European Framework 7 (FP7) funding initiative. The project is a European Union Research Infrastructures project, part of the I3 programme.

Lifemapper, VisTrails and EML

Documented, Re-executable Species Distribution Models

CJ Grady[1], Jim Beach[1], Jeff Cavner[1], Aimee Stewart[1]
[1] University of Kansas
cjgrady@ku.edu, beach@ku.edu, jcavner@ku.edu, astewart@ku.edu

***Abstract*— Lifemapper is an archive of species distribution models as well as a set of web services used to access and create them. We have decided on Ecological Metadata Language for providing metadata for each of our service objects and process metadata for each experiment documenting not only how the process was completed but also how it can be re-executed in the future. Combining this with the clients we have created, we have provided software that can be used to regenerate and re-execute any experiment we have created strictly from the metadata used to describe the inputs, the process, and the outputs. This is especially useful when combined with the VisTrails environment as it gives non-programmers access to powerful tools for scientific experiment generation through a user-friendly graphical interface. Additionally, providing metadata for our service items allows us to track data provenance over time. When this information is added to the documentation of an experiment, a reviewer can see exactly what was done to get from the inputs to the outputs, promoting transparency and reproducible scientific experiments**

Keywords—metadata; software; documentation; reproducability; web services

I. INTRODUCTION

The Lifemapper Project (http://www.lifemapper.org) is a National Science Foundation (NSF, http://www.nsf.gov) funded effort to compile species distributions and computed, predictive range and diversity models. The project is comprised of two primary components. The first is an archive of species distribution models and the second is a collection of web services that access, create, and store data for species distribution modeling and biogeographical experiments. The experiments in the archive are compiled from occurrence data at both the genus and species level acquired from a local cache of species data aggregated by the Global Biodiversity Information Facility (GBIF, http://www.gbif.org) and this data is modeled and projected using scenarios of climate data from WorldClim (http://www.worldclim.org) and the Intergovernmental Panel on Climate Change (IPCC, http://www.ipcc.ch). These inputs are fed to one of several modeling algorithms such as GARP Best Subsets [1] and Bioclimatic Envelope [2] that are available to the openModeller library (http://openmodeller.sourceforge.net). Model outputs are a rule set of habitat suitability parameters and maps indicating the predicted habitat suitability for the organism in question based on the inputs to the algorithm.

The second major component of Lifemapper is its web services. All data and metadata in the Lifemapper system are available from these web services and these services are both RESTful (Representational State Transfer) and Open Geospatial Consortium (OGC, http://www.opengeospatial.org) compliant. Uploading user created content is done using OGC's Web Processing Service (WPS) standard and raster data is retrieved using OGC's Web Coverage Service (WCS) standard for actual data or the Web Mapping Service (WMS) for scaled map images. Metadata about each service item is returned by post pending the desired interface parameter to the end of the REST URL. This metadata includes information about the data that is returned by an OGC service. This can include the inputs to the experiments, keywords, modification time, geospatial and temporal coverage of the data in question, cell size and resolution, or anything else related to the service items.

As an adjunct to our web services, we also provide software clients for users to efficiently access the services through their supported applications. These clients use the published Lifemapper services application programming interface to post and request data and experiments. Our software client integration with VisTrails is especially useful (http://www.vistrails.org). VisTrails (VT) is a scientific workflow management system that allows a user to assemble and document exploratory computational tasks. VisTrails provides a graphical user interface for authoring workflows, parameterizing modules, and for pipelining data through computational steps and output visualizations. A distinguishing feature of VisTrails is its ability to generate comprehensive provenance information or metadata about complete workflows. The result of our LM/VT work is a powerful tool that can be used generate complex experiments while maintaining an easy-to-use user interface.

One of our primary goals is to promote transparency and repeatability in species distribution modeling. For that, we require metadata about the inputs to an experiment. Input metadata includes where the original data can be obtained, any transformations that have been done to the data, etc. Once the input data is thoroughly documented, metadata is produced that records the processes that transform the inputs into the final outputs. Tracking data provenance ensures that we capture all of the manipulations of a data set from start to finish so that we can expose them for evaluation and validation for someone wishing to repeat an experiment [3].

For assembling and archiving Lifemapper workflows, we use Ecological Metadata Language (EML, http://knb.ecoinformatics.org/software/eml/). EML is a metadata specification implemented as a series of XML document types [4]. Our rationale for using EML was that our existing XML metadata fit the schema and could be extended to capture additional information required by the EML standard. Capturing process details is facilitated by the EML specification through the process metadata component in two forms; protocol and method. Allowing processes to be documented descriptively, to explain what was done, and prescriptively, to describe how to do it [5]. This allows us to provide information to replicate the experiment by hand as well as provide instructions that can be used by our clients to automate the experiment replication, a concept we are calling "Executable EML".

At this time, the EML process metadata standard is limited and is primarily a free-form text field that is used to describe the procedural step in a format that is human readable. This is an important first step as it allows for any possible action to be described, however, it falls short with respect to computer-based automation. We acknowledge that not every procedure can be automated because some may be accomplished outside of a computational environment, but we would like to add some capability for automated execution of procedural steps such as web service calls or conditional post processing routines.

II. Progress to Date

Our initial step was to generate EML for all of the data we provide in the Lifemapper archive. Each service provides metadata for each item. Climate layers and species distribution projects are provided as Spatial Rasters and point data is provided as data tables. Additionally, experiments contain the methods used to generate them as well as the protocols to use if the experiment is to be generated again.

Once our services started producing EML, we wrote libraries that could read this metadata. This is our "Executable EML" concept. For this first iteration, only very specific EML can be read and handled correctly, however, from this specifically formatted EML, an entire experiment can be regenerated. The experiment EML includes the methods actually used to generate it, including the software used, as well as the protocol for recreating the experiment. These protocol entries are links to web services that can be called to post data and submit a new experiment using the parameters contained in the document.

With respect to schema, the process metadata we have produced strictly follows the standard EML standard. The text produced is human-readable and notifies the consumer of the web services called as well as what subsection of the document is relevant to that particular step. For example, one of the steps might say something like "Request that the Species Distribution Modeling Experiment be generated by sending an HTTP POST request to http://lifemapper.org/services/experiments with the payload being the content of ref X" and in the document would include a section like Fig. 1.

Our EML reader will parse that text string and then retrieve the referenced subsection and post it to the specified web service. This initial step worked for a proof of concept of how the process metadata could be read and then regenerate the experiment, but it is not general enough for production use and does not include the capability to operate on web services outside of the Lifemapper environment or any new services that we might produce that have different process metadata text strings.

The most visible products we have produced related to EML are our clients, including a Python (http://www.python.org) library and a package of VisTrails modules. The Python library has the capability to read Lifemapper produced EML and produced objects from it that can be used to resubmit the experiment. The Lifemapper VisTrails modules can read Lifemapper produced EML and recreate the workflow used to create the experiment. This workflow will retrieve data for the experiment that is available from a URL. When the workflow is executed, this data will be posted through the Lifemapper REST services and the experiment will be run. The EML may be loaded either in a text box that allows copy and paste or direct entry, or the VisTrails module will go out and retrieve the EML content from a provided URL.

Our efforts to date have created the following architecture represented in the flow diagram in Fig. 2. A Lifemapper experiment starts as a request made to our web services. The accessed web service processes the inputs it receives and submits a job to our processing pipeline. From here, a job is submitted to one of a variety of computational environments depending on the type of job requested, inputs to the job, and other factors so that we can retain optimal performance. If a job can be done quickly and has a small footprint, it may be run locally on one of the main servers. If the job is larger, or would benefit from parallelization, it will be submitted to our compute cluster or broken into smaller pieces and ran in the cloud. After the job has completed, the results are cataloged in our database and relevant files are entered into storage. At this point, an EML document is available through the web service. These metadata documents are not stored in the Lifemapper system and are generated each time they are requested. We do this because, until now, the resulting

```
...
<request reference="X">
   <experiment>
      <algorithmParametersId>XX</algorithmParametersId>
      <occurrenceSetId>YY</occurrenceSetId>
      <modelScenario>ZZ</modelScenario>
      <projectionScenarios>
         <projectionScenario>AA</projectionScenario>
         <projectionScenario>BB</projectionScenario>
      </projectionScenarios>
   </experiment>
</request>
...
```

Figure 1. Sample XML Subsection

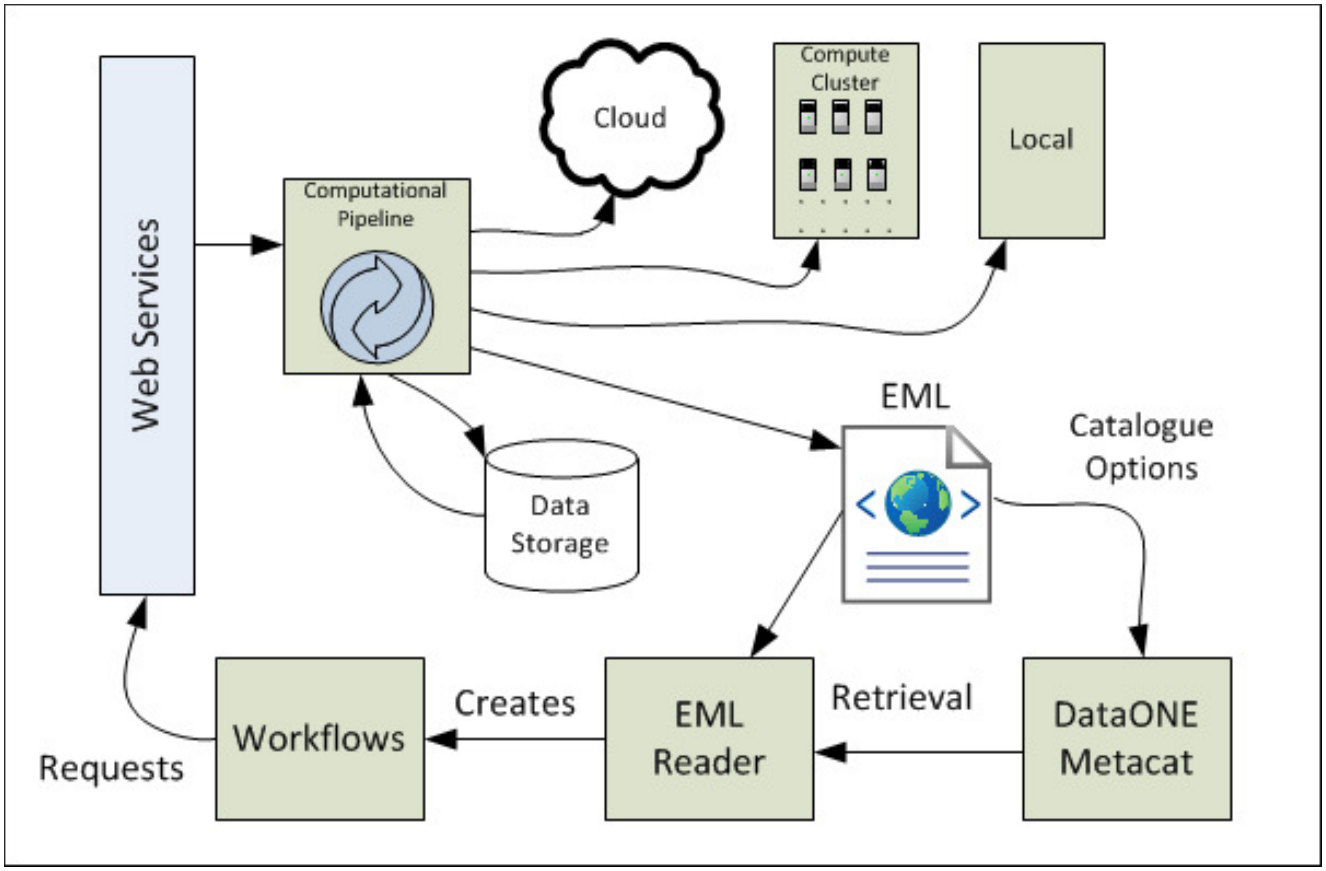


Figure 2. Lifemapper EML Dataflow Diagram

metadata created by each service is predictable and no special operations need to be stored. Once an EML document has been requested it can be sent to some EML storage services. This might be a Metacat (http://knb.ecoinformatics.org/index.jsp) installation or a DataONE (http://www.dataone.org) node. From there, it can be queried and retrieved for future use. An EML document, requested from Lifemapper or retrieved from an EML cataloging service, can then be fed into the EML reader included with the Lifemapper clients which will transform the document into a series of process steps in a workflow. This workflow can be executed through one of our client plug-ins, such as the Lifemapper VisTrails modules, which will make web service calls to the Lifemapper (or other) web services. This cycle gives us "Executable EML" and provides the capability to re-execute and verify scientific experiments.

III. Current Development

We are currently working on extending the EML we produce, and we're continuing to explore the standard to ensure that we are providing the most complete metadata possible. This expansion also includes new services that we are creating. Our goal is to provide EML for every Lifemapper service, and to add mechanisms to ensure that new EML passes validation testing.

Our present efforts also include reading and processing more generic EML. This allows us to handle external data both in our clients, and in our web services. The aim of this is to increase our interoperability with other projects and data sources. This can also allow us to expand our user upload services while at the same time providing a simpler user interface from our clients.

By allowing user uploads via EML, we can get all of the metadata for a climate layer or a collection of occurrence points. This may also be particularly useful if the user's desired data is a large file and available online. The Lifemapper system can just download the data directly, rather than the user downloading the data and then uploading it to our server.

We are currently considering options for EML cataloging and querying. We are initially looking at setting up a Metacat installation for EML storage, searching, and retrieval. We are also exploring the possibility of setting up a portal as part of a collaboration with the University of Oklahoma, Kansas State University, and Oklahoma State University funded by the NSF EPSCoR program (http://www.nsfepscor.ku.edu/). This portal would store event based EML and provide a shared catalog among the institutions. EML will be a primary component allowing for the interoperability of scientific data and processes from multiple science fields.

Since we rely heavily on web services to perform our experiments, it is important that we extend the schema to include metadata about them in order to collaborate with other projects, especially if we expect to generate potentially new cross-cutting science as a combination of the services of each group. Therefore, much of our current efforts have been spent establishing a process metadata schema extension that will capture the procedural steps in a fashion that would be easy to replicate, either by human or automated by a computer. For the current extensions, we have decided to focus on two types of process metadata that we would like to have automated. The first being web service calls and the second being conditional repetition and post processing.

For web service call metadata, we need to capture everything going into the web service as well as everything coming out. This is relatively straight-forward for HTTP GET requests as, often, most of the inputs are included in the URL as query parameters. There are some additional metadata that may be included as HTTP headers as well, and for this iteration, that is all we are including for the requests. The responses need to be documented as well and currently we are capturing metadata about expected HTTP response codes and returned headers as well. HTTP POST requests are a little more complicated as they can include more in the request payload than a GET request. There are a few options available for how this can be accomplished, but we are starting by allowing the metadata to specify an external location of the data via a URL, or the body of the payload can be specified by referencing a section of the metadata document. This can either be raw data encoded into the XML or an XML subsection of the document that will be encoded as the body of the post request.

The second process metadata extension we have focused on is conditional post-processing instructions. At this time, our extension will rely on the response of either a GET or POST request being an XML document and the metadata will include an XPath (http://www.w3.org/TR/xpath/) query to the item to be compared. For instance, our species distribution modeling service produces a metadata document that includes a status variable. A conditional processing step in this case would be set to repeat the request for this document until the status indicates that the experiment had completed, at which time the next procedural step would take place. Combined, these two new process metadata extensions allow us to prescribe how an experiment can be generated in a way that will allow an EML reader to automate the re-execution of the process. Additionally, these extensions are general enough that they can be used with web service calls outside of the Lifemapper system.

IV. Future Work

In the future, we would like to introspect the EML specification XML schema definition file to create Python objects representing each element. This approach provides multiple benefits for creating and parsing EML. New EML will be easy to validate if type checking is added to these objects. Uploaded EML documents will be validated and quickly transformed into a tree structure that will provide simple access to requested data.

We would like to expand our clients to publish EML to any server requested. This will include any EML catalog associated with Lifemapper, any DataONE node, or any other server that takes EML from a HTTP POST request. We will also expand the clients to generate EML for anything produced in the client. This improved interface will work similar to the way Morpho (http://knb.ecoinformatics.org/morphoportal.jsp) works and will allow users to document their newly created experiments.

The next steps for our process metadata extensions are to continue to research web service call metadata in pursuit of a standard metadata format that we can leverage if one emerges. As our process metadata is in it's infancy stage and currently highly fluid, we would like to use a standard created by some governing body if possible for HTTP requests and responses as it would likely be more complete than something we are able to generate.

Our conditional processing metadata will be expanded as well. Some of these additions might be adding support for JSON condition analysis and possibly using XQuery (http://www.w3.org/TR/xquery/) to for evaluation. Using a standard such as XQuery will give us greater flexibility when creating a client to read and evaluate our process metadata as well as provide a standard for other developers to use when utilizing these documents.

Lifemapper's use of EML is helping us accomplish our goals of queryability, self-standing metadata, interoperability, and repeatable science. By using EML with our own and external portals, users can search for our data that is related to their interests. Providing EML documents with all information needed to recreate an experiment allows users to recreate and verify experiment results without using the Lifemapper system if they so choose. They are able to acquire all of the data used and know how to process it from the metadata as well as the actual procedure used to create the experiment. They can also use our clients and "Executable EML" to read all of the metadata about an experiment, get all necessary data, and then re-execute the experiment automatically. Overall, we have become a more viable option for collaboration with other projects, expanded our user-base, and are promoting transparent and repeatable science.

Acknowledgment

The Lifemapper project is funded by NSF grants EPS 0919443, DRL 0918590, BIO/EF 0851290, and OCI 0753336.

References

[1] Anderson, R. P., D. Lew, and A. T. Peterson. "Evaluating Predictive Models of Species' Distributions: Criteria for Selecting Optimal Models," Ecological Modelling, vol. 162, pp. 211-232, 2003.

[2] Nix, H. A. "A Biogeographic Analysis of Australian Elapid Snakes," Atlas of Elapid Snakes of Australia, pp. 4-15, 1986.

[3] Freire, J., D. Koop, and L. Moreau (Eds.). "The Open Provenance Model: An Overview," IPAW, LNCS 5272, pp. 323-326, 2008.

[4] Ecological Metadata Language (EML) http://knb.ecoinformatics.org/software/eml/

[5] Ellison, A. M., L. J. Osterweil, L. Clarke, J. L. Hadley, A. Wise, E. Boose, D. R. Foster, A. Hanson, D. Jensen, P. Kuzeja, E. Riseman, and H. Schultz. “Analytic Webs Support the Synthesis of Ecological Data Sets,” Ecology, 87 (6), pp. 1345-1358, 2006.

Moving from Custom Scripts with Extensive Instructions to a Workflow System: Use of the Kepler Workflow Engine in Environmental Information Management

Corinna Gries[1], John H. Porter[2]
[1] Center for Limnology, University of Wisconsin
[2] Department of Environmental Sciences, University of Virginia
cgries@wisc.edu, jporter@virginia.edu

***Abstract*— Here we discuss the applicability of the Kepler workflow system for basic environmental data management. Examples for its current use by two Long-Term Ecological Research sites are given in the areas of basic table manipulations, managing streaming sensor data, and quality control routines involving complex R scripts. Overall we find Kepler a very good tool for the task and particularly well suited for a community of practice in which specific knowledge transfer may reduce the otherwise steep learning curve. Employing a powerful and flexible platform like Kepler by such a community can provide for extensive exchange of expertise and widespread reuse of workflows and specifically developed actors.**

Keywords—Kepler; workflow; environmental information management; data curation

I. INTRODUCTION

The benefits of using scientific workflow systems are generally discussed in the context of data analysis and modeling by scientists involving distributed computations and large amounts of data [1] [2]. Workflow systems can abstract data access, manipulations, analysis, and visualization, as well as parameterize and run models, stage data and schedule remote jobs. This not only results in an automation of multi-step processes, but also documents data provenance and model parameterization as well as workflow evolution [3]. Other benefits include the reuse of workflow components and publication and exchange of entire workflows. Most workflow engines (e.g., Kepler [4], Taverna [5], Triana [6], VisTrails [7]) allow encapsulation and chaining of different analytical tools. These workflow systems have in common that they provide a graphical user interface for designing workflows but may differ widely in the computational approaches used for their design and processing of workflows [8,9]. Here we focus on Kepler, which is distinguished from other workflow systems by its inclusion of iteration, user-defined workflow scheduling, use of both abstract and concrete model structures and availability of both task and workflow-level fault tolerance [9].

We chose Kepler for our information management tasks because of its strong support of general database interactions and some specialty actors (e.g. EML [10], DataTurbine [11]). The Kepler workflow engine provides an intuitive graphical user interface and a wide range of workflow components, called 'actors'. Actors encapsulate generic functionality for data input, conversions and calculations, output, general purpose functionality, workflow control, and specific functionality for accessing several analytical packages. These actors may be dragged onto the workflow canvas and then connected by their input and output 'ports' while a 'director' controls the data flow. Each director represents a different computing model and the most commonly used ones come with the Kepler installation. In addition to the actors and directors provided with the Kepler installation, a searchable repository for custom actors is available (for a more detailed description of Kepler see [4]). User input may be specified via 'parameters' and annotations may be entered for every step or the overall workflow. Although the graphical user interface is very intuitive, the screen can become cluttered in complex workflows. This can be alleviated by gathering groups of actors, representing related functionalities, into 'compound' actors. This not only makes the workflow more readable, but also facilitates the reuse of these compound actors as individual components in new workflows.

Although widely used in data analyses, the benefits of using scientific workflow systems extend beyond analyses, modeling and large-scale science. The open source workflow system Kepler has utility in basic ecological data management involving data manipulations and quality control procedures routinely conducted during the data-curation process before data are ready for analysis, synthesis, and modeling. The applications presented here pertain to long-term environmental

monitoring, where the same data are collected monthly, annually, or are streaming in real-time.

Before the use of a workflow system, a combination of esoteric scripts and proprietary software would be used to perform conversions. Specifically, Perl, Fortran, PHP scripts/programs, the Oracle-specific SQL language, PL/SQL, and Oracle triggers were used in addition to MS ACCESS and MS Excel. Data were moved between Windows and Linux systems and converted to different proprietary formats to accomplish most of the tasks. Extensive natural language documentation of the procedures, the location and idiosyncratic requirements of custom scripts were necessary to be able to repeatedly perform all involved steps, especially if the data handling occurred once a year or was performed by different people. Kepler workflows not only save time by automating the execution of these numerous steps, but also provide graphical and text documentation in an easily-interpretable standardized format. Although a formal performance analysis would be impossible to conduct due to the large number of manual steps involved prior to employing Kepler, we can assert that the time necessary has decreased from several tens of minutes to seconds.

The capability of documenting transformation steps and data provenance is particularly important when harmonizing data from different sources into an advanced data product for synthesis and modeling. Frequently, similar data are collected with different methods, at different time intervals, and reported in different units and data models. Invariably some aggregation and conversion is necessary before they can be provided as an advanced, value added data product to be used in analysis or modeling and documenting data provenance is paramount.

The here developed Kepler workflows are currently not publicly accessible, but are available upon request. The website 'myExperiment' (http://www.myexperiment.org/) does not seem appropriate for this kind of workflows. However, they may well be submitted to a more appropriate community website in the future.

II. EXAMPLES OF USING KEPLER

A. Kepler for basic data management applications

At the North Temperate Lakes Long-Term Ecological Research (NTL LTER) project many long term monitoring data are collected manually on a monthly or annual basis. Data collection often involves many steps spread out over long periods of time. For example, sample jars would be weighed, labeled and stored in the winter, samples taken some time during the summer field season along with related measurements and then analyzed still later in the lab. Different data entry systems were developed and tested, For example, data entry applications for hand-held devices (PDA) in the field, online web forms for data entry, and data entered into standardized Excel spreadsheets. We found that online forms were generally unacceptable to data entry personnel, and that good data curation and archiving required Excel spreadsheets and PDA output files to be quality controlled and parsed into a database. Transforming data from forms suitable for data entry into forms suitable for archiving and analysis frequently involved transposing or otherwise changing the structure of the tables. Quality control procedures include spell checking species names, verifying collection dates and sites, range checking measurements, eliminating duplicate entries, assuring consistent data types in columns (i.e., avoiding comments in a data column), and evaluating data completeness. Many of these steps had been conducted manually and by taking advantage of functionality in proprietary applications, for example, manipulating the data structure in Excel, using custom and database engine-specific scripts for parsing and loading into a database, stored procedures and *ad hoc* SQL queries for verifying collection dates and spell checking species against authority tables, and using database triggers for range checks. Because the spreadsheets were standardized and didn't change from year-to-year and the necessary steps for data processing were well-documented, we were able to directly translate them into Kepler workflows employing only standard actors. These well-annotated workflows now provide, not only complete automation of these multi-step procedures, but they also contain all necessary documentation of the steps taken to manipulate the data. Additionally, these scripts are database-engine agnostic, encapsulate some of the more complicated custom scripts developed earlier, and are independent of proprietary software and data formats.

The simple Kepler workflow depicted in fig. 1 contains all information to transform a comma delimited text file containing the water levels for 39 ground water wells, which are collected monthly. Any given raw file usually has several months worth of data. The first column is the sample date, the second a correction factor followed by 39 columns, one for each well. This format is efficient for data entry, but poor for archiving and analysis. Therefore, this table needs to be transposed into the database format of: sample date, well ID, water level, and flag. The date has to be reformatted, the water level calculated from meters well depth to actual elevation based on the known well head elevation and the correction factor, and a mix of data types in the columns where dry or frozen wells are noted along with the meters depth, needs to be converted to an appropriately flagged missing value. Prior to using a Kepler workflow, the approach included many formatting steps and moving of data within Excel, specifically copying of the well elevation from year-to-year, copying the formula, and a custom parsing and uploading script. The natural language step-by-step instructions required approximately 1.5 pages. To simplify the graphical display, this Kepler workflow contains two compound actors, one for formatting the date and the other for calculating the ground water level, transposing the record and inserting appropriate flags (iterate Over Array). If desired, these compound actors can be opened to display the processing details, just as for the overall workflow.

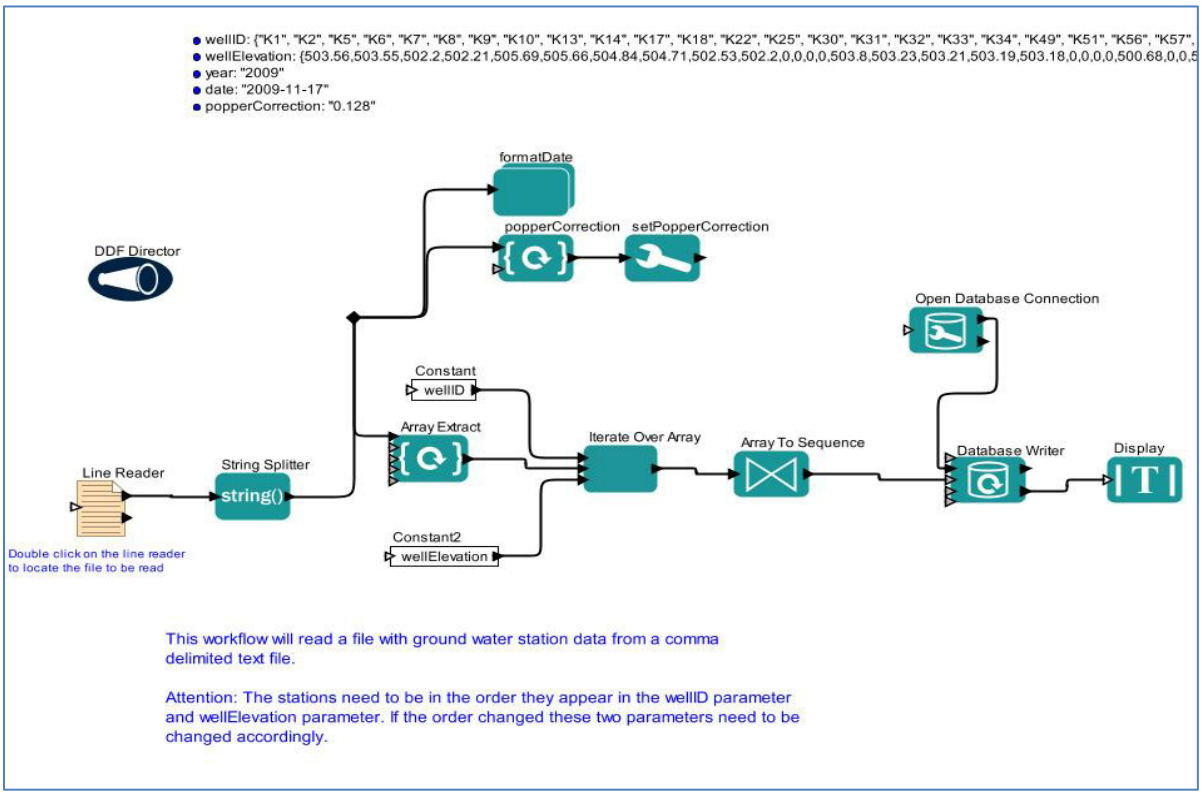


Figure 1. Kepler workflow for parsing comma delimited text file with ground water level data into a database

B. Kepler for Managing Sensor Data Streams

Following the lead of the REAP project (Real-time Environment for Analytical Processing [12]) [13] we experimented with establishing workflows to quality control, monitor, and parse sensor streams into final database storage. Data streams from seven lake buoys, each holding up to 20 individual sensors currently are read into a DataTurbine server at North Temperate Lakes LTER. Originally, these data streams were parsed by a DataTurbine off-ramp into a temporary database table and database triggers would apply range checks, parse the data into final tables and calculate hourly and daily aggregates. Employing the DataTurbine actor in Kepler this extra step and database-specific trigger may be bypassed adding the option of monitoring the data streams in real time. We are expecting major improvement to our simple workflows by the developments in the REAP project.

C. Kepler for Quality Assurance and Control

At the Virginia Coast Reserve Long-Term Ecological Research (VCR LTER) project, Kepler is used to help produce statistical displays and graphics to aid in quality assurance and control activities (fig. 2). These processes make extensive use of the link between Kepler, the Ecological Metadata Lanaguage (EML [10]) and the R statistical language. A typical workflow ingests an EML Metadata document containing metadata for one or more tables in a dataset. XML actors in Kepler transform information in EML into an R program capable of reading data tables in the dataset. Additional Kepler actors are used to edit that program to incorporate specific information needed for its operation, such as the location of data files on the local system. The R program is then executed to ingest the data and save it as an R workspace. This method is used in preference to using the built-in EML actor because that actor can become overloaded when large (>10 MB) data input files are used. An additional R program, in a separate actor, can be customized to produce graphical displays that aid in spotting problems with data, such as sensor drift, can also be incorporated into the workflow. The workflow can either be run via the graphical user interface, or as a "batch" job for periodically updating displays.

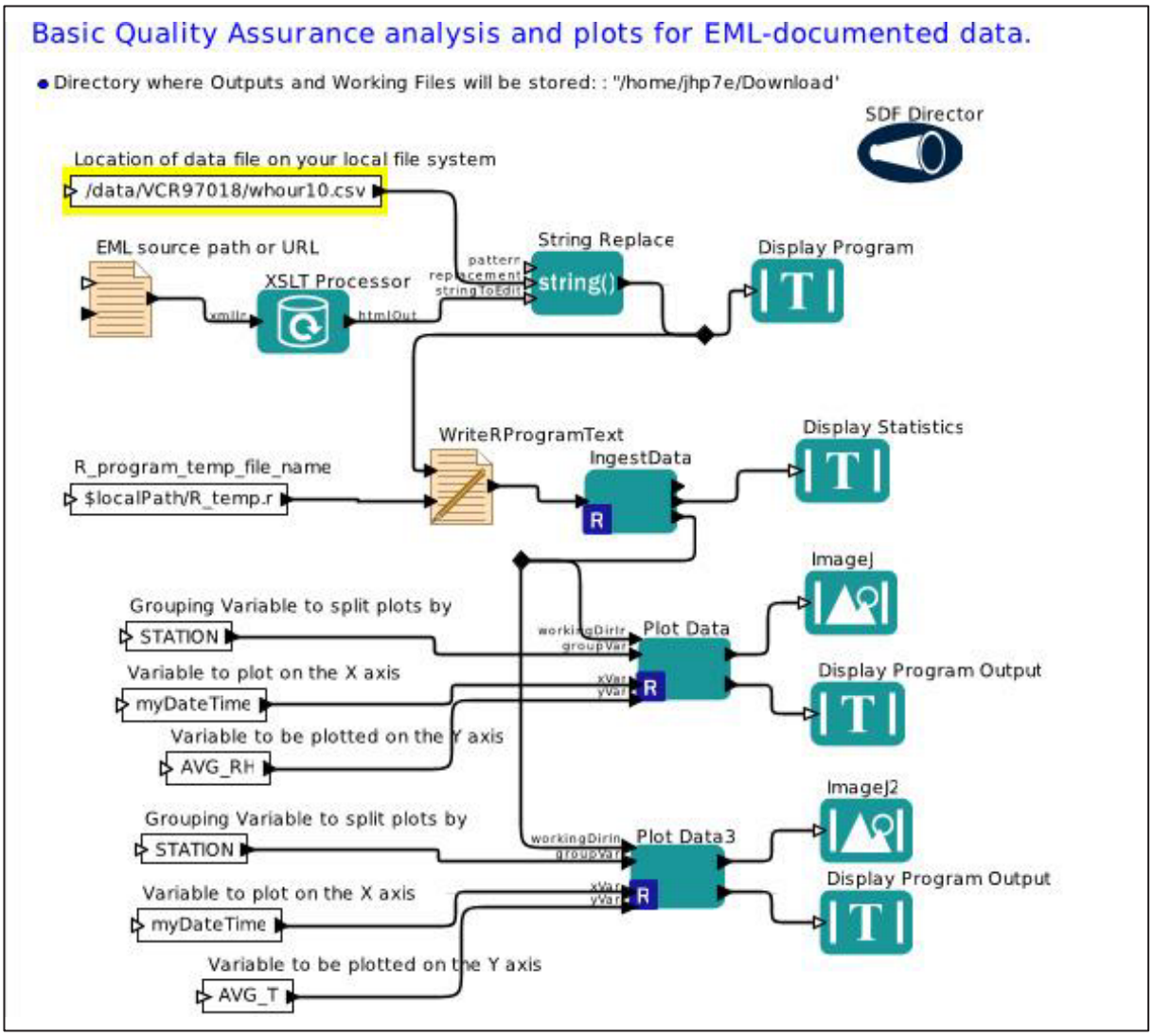


Figure 2: Kepler workflow for displaying statistical QA/QC analysis

III. Advantages and Disadvantages of Using Kepler

A. *Improving documentation and communication*

As shown by the examples, Kepler workflows provide substantially improved mechanisms for documenting and communicating data management procedures. A single Kepler workflow can replace large numbers of esoteric scripts (which require extensive documentation, or may require location-specific configuration). The graphical display makes it easy to see how processing steps are linked together, and on-screen annotations can provide what few instructions are needed to run a workflow. The examples above illustrate how many distinct operations and several pages of documentation could be encapsulated in a single Kepler workflow. Kepler also has advantages for communication of procedures between researchers. A single Kepler workflow can be much more easily transported than large numbers of scripts, and will run properly on different operating systems. We would like to emphasize again, however, that we are dealing with consistent input file formats and only minor changes in the workflows are necessary at runtime.

B. *Flexibility*

Kepler provides a good framework for combining capabilities from several different software packages. It supports a wide array of 'actors', including support for the R statistical language, relational database access, the Python language, the ImageJ graphics package, web services, processing of XML documents, remote processing on Linux and Unix computers, automatic ingestion of datasets documented using Ecological Metadata Language and even proprietary software such as Matlab (although a separate Matlab installation is required). Storing location-specific information, such as the file paths used on specific systems, can be easily accommodated by defining parameters, essentially variables that can be accessed within Kepler actors. This makes

it much easier to transfer workflows between individual computers or even operating systems.

C. Extensibility and Reuse

Clearly an experienced PHP or Python programmer could similarly script the workflows described here without relying on manual execution of separate steps. However, Kepler provides basic capabilities listed above as actors which can be reused in different workflows and provides a graphical user interface for doing so. Therefore, designing a simple workflow for managing data can be put together with limited programming knowledge. Additionally, it is fairly simple to wrap custom code written in Python and Perl or single Java statements as well as scripts in R or Matlab in Kepler actors. This does not require developing a custom actor from scratch. Although we have not had the need, programming a new actor within the Kepler framework will allow reuse of this effort in other workflows and by other users and a central actor repository where custom actors may be exchanged is available and can be accessed and searched from within Kepler. Using a workflow system rather than custom scripts has of course many advantages in large and distributed science applications [14] which may not be applicable to the small scale day-to-day environmental information management requirements. However, the re-usability of components in many small and very similar, yet slightly varying, workflows in a single environmental information management system makes it well worth the effort of learning Kepler. Additionally, the aspects of extensibility and reuse seem particularly well suited to a distributed community of practice like the LTER information management community in which similar tasks are performed across many different systems most of which currently are custom coded.

D. Caveats

Of course, no such system is perfect and we have found a few limitations of using Kepler strictly for data management application as well as more general reservations. Kepler is very powerful and flexible allowing for many different applications and approaches. The flip side of this is that the learning curve is fairly steep. Although simple workflows can be written almost instantaneously in Kepler by following the quick start guide [4], we found that a week or two of intense trial and error and studying the documentation on the Kepler website [4] were necessary before routinely designing more complex workflows. Without any familiarity with programming or data structures it can be hard to design workflows that link together different components (e.g., R-scripts with Python programs) and to use Kepler efficiently. Kepler maintains the strong data typing from the Java language, which is good because it helps make workflows robust, but it can be difficult for the beginner to successfully move data from actor to actor. Additionally, although conversion actors are provided, date and time handling is still as cumbersome (or well controlled) as in Java.

As mentioned above, large datasets may be handled inefficiently in certain actors and in exceptional circumstances even cause Java crashes. Additionally, documentation of individual actors is often minimal to absent or sometimes highly technical. The same is true for error messages. For instance, R scripts need to be thoroughly debugged prior to running in Kepler because error messages from R are lost if the program crashes (although the try() and geterrmessage() functions in R can be used as an alternative to debugging outside Kepler).

Most scientific workflow systems are built on the data-flow model rather than the control model implemented in business-oriented workflow systems. Kepler offers some more flexibility by providing different 'directors' which support a variety of computational models including data flow and control flow [8,9]. However, in contrast to many data-driven scientific analysis workflows some of our more complex data management workflows have some components of control flow where scheduling may be complex and not data flow dependent (e.g. database interactions involving table creation before data already in the pipeline can be streamed into the new table). We found it very difficult to implement these types of workflows. This may of course be due to our still limited understanding of Kepler's full capacities.

IV. Conclusions

Kepler has reached a level of maturity where it can be reliably deployed in production environments for data management. It continues to be improved and environmental data management specific actors are being developed by currently funded projects. Using a flexible and powerful platform like Kepler to streamline general data management procedures seems particularly appropriate for a distributed community of practice. Specific knowledge transfer can minimize the learning curve and custom developments may benefit the broader community.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grants No. 0621014 (VCR LTER) and #DEB-9632853 (NTL LTER). Any opinions, findings, conclusions, or recommendations expressed in the material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludäscher, and S. Mock, "Kepler: An Extensible System for Design and Execution of Scientific Workflows." Proceedings of the 16th International Conference on Scientific and Statistical Database Management, IEEE Computer Society 21-23 June 2004.

[2] Y. Gil, E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moraeay, and J. Myers, "Examining the challenges of scientific workflows", Computer, vol.40 (12), pp. 24-32, 2007

[3] B. Ludäscher, I. Altintas, S. Bowers, J. Cummings, T. Critchlow, E. Deelman, D. D. Roure, J. Freire, C. Goble, M. Jones, S. Klasky, T. McPhillips, N. Podhorszki, C. Silva, I. Taylor, and M. Vouk. "Scientific Process Automation and Workflow Management", In Scientific Data Management: Challenges, Technology, and

Deployment, Computational Science Series, A. Shoshani and D. Rotem, eds. chapter 13, 2009

[4] The Kepler Project https://kepler-project.org/

[5] Taverna Workflow Management System http://www.taverna.org.uk/

[6] Triana, The open source problem solving environment http://www.trianacode.org/

[7] VisTrails http://www.vistrails.org

[8] E. Deelman, D. Gannon, M. Shields, and I. Taylor, "Workflows and e-Science: An overview of workflow system features and capabilities", Future Generation Computer Systems 25, pp. 528 - 540, 2009

[9] J. Yu and R. Buyya. 2005. A taxonomy of scientific workflow systems for grid computing. SIGMOD Record, Vol. 34, No. 3, Sept. 2005 pp. 44-49.

[10] Ecological Metadata Language http://knb.ecoinformatics.org/software/eml/

[11] DataTurbine http://www.dataturbine.org/

[12] REAP Project (Real-time Environment for Analytical Processing) http://reap.ecoinformatics.org/

[13] D. Barseghian, I. Altintas, M. B. Jones, D. Crawl, N. Potter, J. Gallagher, P. Cornillon, M. Schildhauer, E. T. Borer, E. W. Seabloom, and P. R. Hosseini, "Workflows and extensions to the Kepler scientific workflow system to support environmental sensor data access and analysis" Ecological Informatics 5, pp. 42-50, 2010

[14] B. Ludäscher, M. Weske, T. McPhillips, S. Bowers. "Scientific workflows: Business as usual?" In *7th Intl. Conf. on Business Process Management* (BPM), Ulm, Germany, 2009

A Cyber-Infrastructure for a Virtual Observatory and Ecological Informatics System -VOEIS

Clemente Izurieta[1], Sean Cleveland[1], Ivan Judson[1], Pol Llovet[1], Geoffrey Poole[1], Brian McGlynn[1], Lucy Marshall[1], Wyatt Cross[1], Gwen Jacobs[1], Barbara Kucera[2], David White [3], F. Richard Hauer[4], Jack Stanford[4]

[1]Montana State University, Bozeman MT 59717

[2]University of Kentucky, Lexington, KY 40506

[3]Murray State University, Murray, KY 42071

[4]Flathead Lake Biological Station, Division of Biological Sciences, University of Montana, Polson, MT 59860

clemente.izurieta@cs.montana.edu

***Abstract*— The Virtual Observatory and Ecological Informatics System (VOEIS) provides a framework for data acquisition, analysis, model integration, and display of data products from completed workflows including geospatially explicit models, graphs from statistical analyses, and GIS displays of classified ecological attributes on the landscape. VOEIS is intended to complement the capabilities of the Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI) Hydrologic Information System (HIS) by providing sound data and metadata management capabilities for field observations and analytical lab actions. Functionality provided by VOEIS is supported by a Field Data Model (FDM) that enhances the limited geospatial capabilities of CUAHSI's Observations Data Model (ODM). Access to VOEIS data and metadata is also made accessible via programmatic APIs which facilitates integration with other service oriented "*e-Science*" architectures and distributed frameworks.**

Keywords—framework; cyber infrastructure; data and meta-data management

I. INTRODUCTION

CUAHSI's Hydrological Information System (HIS) is an internet-based system that supports the distribution of hydrologic data. CUAHSI's HIS "is comprised of hydrologic databases and servers connected through web services as well as software for data publication, discovery and access." [1,23] Though HIS provides exceptional server side support, data entry and quality control client tools, HIS presumes that individual research labs posses sound internal data management practices, doesn't provide tools for managing metadata about field and analytical lab actions, and has a limited data model for geospatial reference. CUAHSI's Observations Data Model (ODM) [6] is founded upon an information model for observations at stationary points. This model is insufficient to characterize complex spatio-temporal relationships that arise under circumstances where hierarchical and dynamic sampling locations occur. VOEIS is an integrated sensor and ecological informatics system that complements CUAHSI's HIS capabilities by supporting all-encompassing workflows; from the collection of streaming senor data to the application of those data in simulation models and visualizations. VOEIS facilitates the management of data and science metadata within individual research labs, solves the problem of the static geospatial data model, and interfaces with HIS to allow labs to share some or all data via the HIS protocols.

The VOEIS infrastructure is designed to extend the functionality and knowledge representation capabilities of CUAHSI HIS by providing necessary interfaces, software components, and a complementary Field Data Model (FDM) schema [18] that captures data processed in the lab or collected by scientists in the field.

VOEIS has three basic research elements: 1) the development and deployment of sensor networks which requires the cyber-infrastructure enhancement of hardware at two field hubs (FLBS and HBS described in section III B); 2) the development and deployment of an informatics system to manage and serve hydrological and meteorological data and metadata, and to interface with CUAHSI's HIS and ODM; and 3) the development and usage of protocols and APIs to interface with partnering technologies (i.e., WaterML [27]).

II. BACKGROUND

The challenges of managing scientific data are significant, and over the years they have typically fallen in the hands of investigators. There exist significant obstacles in workflows supported by cyber-infrastructures; from operation and field deployment of sensors – to data streams – to data management – to data analysis – to the use of integration tools. These multifaceted obstacles involve hardware, middleware and software. However, significant work and progress has been made to tackle the challenges of managing these workflows, discovering data, storing data, and publishing scientific data in architectures that are conducive to ease-of-use, dissemination, documentation and research for scientists. PIs, researchers, managers, and scientists alike need the ability to easily access (and possibly integrate) information that is housed in distinct geographical and distributed sites. Additionally, such information is very likely to be stored in different formats and disseminated using a diverse range of communication protocols. The Tupelo middleware [4] developed at the National Center for Supercomputing (NCSA) and the

University of Illinois is an open source semantic content management framework (middleware technology stack) designed to manage e-Science projects. This is an all-purpose solution whose goal is to manage information from a broad range of sources and to provide functionality that supports data management, provenance, workflows, people, and temporal and geospatial relationships. Similarly, the NSF sponsored Data Observation Network for Earth (DataOne) [2] project has undertaken the task of developing a distributed framework and cyber-infrastructure to support the needs of the e-Science community. DataOne tackles the data integration problem by developing standards based technology to support all encompassing biological sciences domains, i.e., hydrology, ecology, atmospherical, oceanographic, etc. DataOne uses a service oriented architecture centered on collaborating nodes that maintain registries of available data and their addresses. To participate in DataOne, a researcher may choose to create a member node and implement its associated interfaces. Additionally, participating member nodes may choose to implement a full set of APIs that allows the member node to also accept data from other participating nodes. This allows clients to access and share information, which reduces the possibility of data loss and allows researchers to aggregate and analyze data from many sources. Clients interact with member nodes by using anyone of many services available through an investigator toolkit.

VOEIS, in contrast, is focused on hydrological and meteorological data only. VOEIS enhances and expands the information made available by CUAHSI's ODM through its associated HIS server. A VOEIS server can be integrated into any service oriented framework. For example, you can turn a VOEIS server into a DataOne member node by implementing the desired interfaces, registering it with a coordinating node, and mapping content schemas. VOEIS provides a programmatic RESTful API that can easily interface (via a façade for example) with other APIs, and our underling evolvable schema technology design [8] allows for the flexibility to represent content in other formats and provides a mechanism that supports dynamic changes to schemas. Participation in the greater e-Science community is a VOEIS goal, and the technical aspects of interfacing in the DataOne network and the Tupleo technology stack are currently being assessed. In the next section we describe the VOEIS architecture and all services made available to potential client nodes.

III. Architecture and Functionality

This section contains abridged structural and behavioral details of VOEIS. A high level description of the software architecture, related data collection functions and user interface provide an overview of VOEIS functionality.

A. *Architecture – High Level Overview*

The VOEIS Data Hub is an open source data management and publication software stack designed to store and organize hydrological, water quality, water chemistry, and meteorological data. Investigators can organize data into projects and create geospatial sites that are associated with temporal-tagged observations, sample readings, and sensor measurement data. VOEIS is designed to support research lab style data management, data collaboration and data publication through its own web presence and through the CUAHSI HIS services.

VOEIS (see Figure 1) is implemented using the Yogo Framework [9] with Ruby on Rails [20] to take advantage of the data management tools providing flexible schema management, RQL API, Role Based Access Controls (RBAC), versioning and support of multiple database back-ends making it platform independent. Currently the system uses PostgreSQL [19] as the backend storage system. Data processing has been optimized for PostgreSQL; however generic implementations can make use of any DataMapper ORM supported backend such as MySQL [13], SQLite3 [22], Persevere [17], MongoDB [12], etc. Yogo is open source software and is available for download [26].

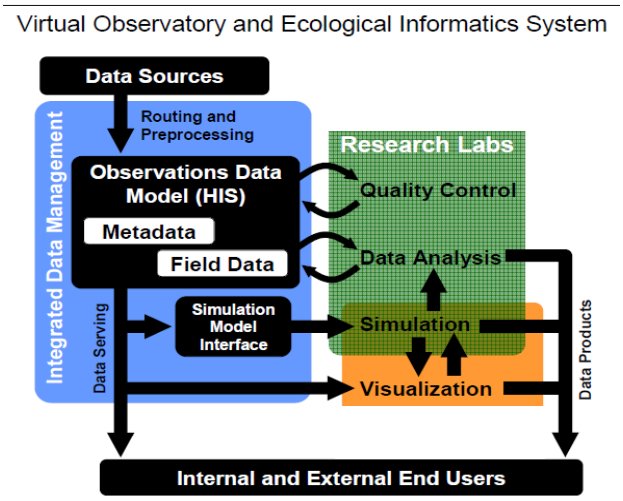


Figure 1. High level architectural view of VOEIS

B. *Sensor Data Collection*

The means of collecting ecological data include deployed lake, river and meteorological sensor systems. Moving vast quantities of real-time data from deployed sensor systems requires innovative wireless, satellite, cell, and/or combinations of these systems. The VOEIS Data Hub currently collects streaming data from three different sources. The Big Sky sensor array consists of four stations deployed in distinct stream localities and one weather station used to collect raw data transmitted from each station via high radio frequencies. All stations are equipped with Campbell Scientific CR1000 data loggers that store hydrological and meteorological data. The other two data sources are deployments in lakes. Both Flathead Lake Biological Station (FLBS) located in Montana and Hancock Biological Station (HBS) located in Kentucky import meteorological data and lake buoy hydrology and water quality measurements into the system. VOEIS currently supports parsing CSV files from data loggers and text based data from samples organized as time-series. Both biological stations are constantly inundated with requests for data from researchers and also the public. The data managers of both stations are busy handling operations for importing and

curating data as well as creating reports and archives, and are thus challenged and confronted with meeting the expanding demands for data from their internal clients and from a public that has become aware of the usefulness of the data for fishing and boating purposes. VOEIS aims to alleviate some of the demand on these individuals by allowing that both internal and external clients have appropriate access to the data and are able to search and acquire it in their own time with little intervention.

C. *Data and Meta Data Management*

VOEIS is designed to manage information, and the science and administrative metadata required to make the data useful to other data consumers. VOEIS is able to capture the current data and metadata that CUAHSI HIS ODM 1.1 is designed to capture plus additional data types and additional metadata significant to the research and lab data management processes.

VOEIS uses an evolvable schema technology and data paradigm in order to easily support the ability of scientists to modify data models quickly [8]. Unlike relational technologies that require significant design work *a-priori*, an evolvable schema supports schema alterations *during runtime* that are necessary to support new functionality.

D. *Field Data Model (FDM)*

The goal of FDM is to provide a complementary schema that characterizes complex spatio-temporal relationships that cannot be realized by ODM. FDM captures the structural relationships necessary to augment ODM. It is not the intention of FDM to accommodate for the modeling of information fluxing through an environment. A simulation modeling interface, depicted in Figure 1 will be provided to support the ability to interface with VOEIS in order to generate said simulations. Significant work to develop simulation, hydrological modeling frameworks has been done by [5,7,25].

FDM is a significant contribution (developed over two decades) to evolving a data schema that allows functional integration of data from field observations, analytical labs, and data loggers whose format can be efficiently queried regardless of data source. The FDM can be broken into five basic components (shown in Figure 2), and a resulting unified database of results.

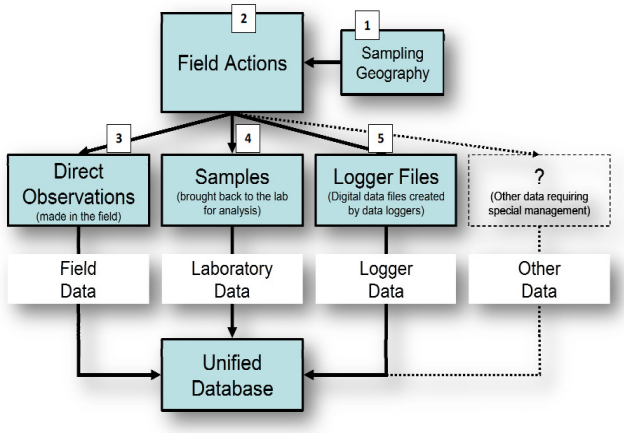


Figure 2. Schematic of general data flow for integrating data from field observations, lab analysis of samples, and results from data loggers

The components address the management of 1) geographic (meta)data describing the locations of study sites and sampling locations; 2) (meta)data describing actions that occur in the field (direct observations, sample collection, and logger deployment/retrieval) at study sites and sampling locations; 3) (meta)data about direct field observations; 4) (meta)data describing and tracking laboratory analyses that generate lab data; and 5) (meta)data describing logger deployments, retrievals, and resulting raw data files.

In VOEIS, we extend ODM with the FDM to provide investigators with the most flexible and robust solution that supports the ability to store, manage and publish data. Figure 3 is a simplified version of the structural UML [24] class diagram that represents the schema of FDM. There are four types of objects: 1) administrative objects represent the set of classes necessary to identify projects, their members, permissions (not shown) and a TupleID used to associate a project with a campaign and visit tuple; 2) action objects represent various field actions, each of which can be associated with data collected for said activity; 3) temporal objects which represent the time characteristics associated with actions; and 4) spatial objects, which represent the geospatial information associated with the actions.

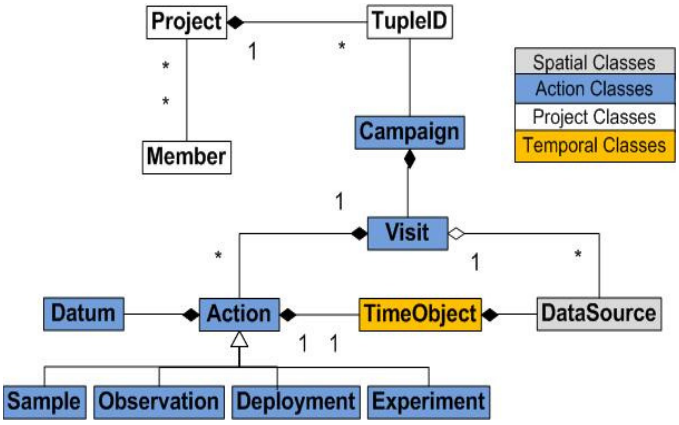


Figure 3. Simplified UML class diagram of the Field Data Model (FDM)

The modular design of VOEIS is intended to allow integration of other components and data types that originate as a function of field work, but require different data management pathways. To illustrate why an FDM is necessary, consider that observations are made, samples are taken, and sensors are deployed at specific points in three dimensional spaces. In order to catalog and track the location of field "actions" (e.g., observations, samples, or deployments) in a database, the action occurs at a "place" (e.g., monitoring water quality at a conceptual location such as the "mouth of a river"), but that the geographic location of virtually any conceptual "place" may change over time (e.g., lateral erosion of a river bank during high flow can cause the location of the "river's mouth" to migrate to a new geographic location). FDM allows field actions to occurs at "places" (DataSource in Figure 3), while "places" can be associated with multiple spatial locations for different periods of time (TimeObject in Figure3).

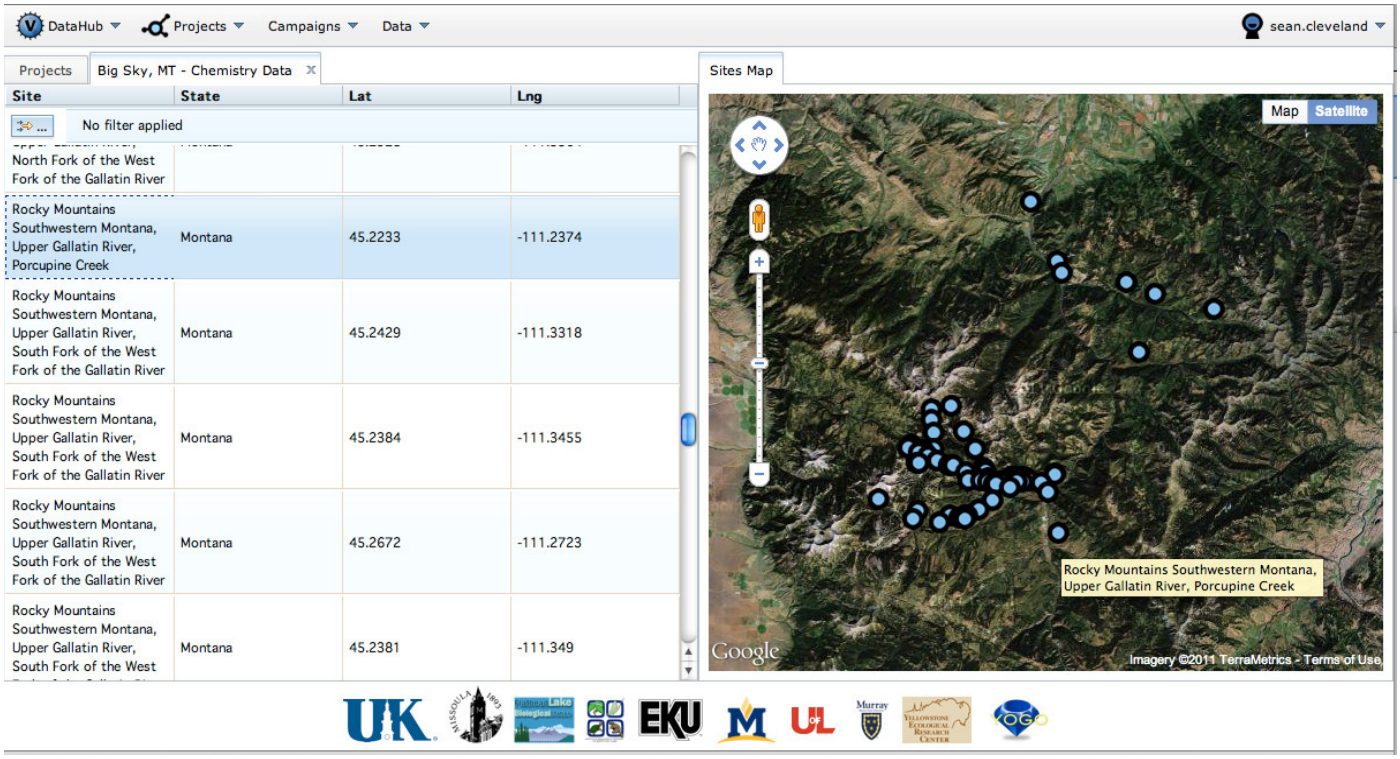


Figure 4. VOEIS UI interface for project browsing

E. Graphical User Interface

The VOEIS UI is served through basic html and Dojo Toolkit JavaScript widgets [3] to support multiple browsers. The VOEIS UI is specifically designed for investigators. A contextual inquiry process along with numerous requirement validation meetings were carried out before settling on a project centric UI. VOEIS allows investigators to display managed projects with options to upload, browse, search and download data. The current UI supports simple upload of workflows for logger and sample/chemistry data within a project with options that allow saving the resulting parsing instructions for re-use. Simple data graphs are supported for displaying time-series data. The VOEIS UI provides simple views of project data that support the end-to-end public drill down of data products. In Figure 4 we display a simple representative example of the UI.

F. Programming Interface and Data Sharing

The VOEIS Data Hub is designed to offer developers programmatic access to data via RESTful APIs. The APIs provide significant flexibility for users to create personalized tools and views to interact with the data stored in VOEIS. Access to API methods is role based and is implemented using API keys that are linked to user accounts. Thus, API access is managed in the same manner that regular user access is handled. Further, the implementation of a Resource Query Language (RQL) REST interface allows dynamic data querying API functionality resembling SQL database queries without the security risks and complication inherent in exposing an SQL interface for a complex database schema.

Once data are in common formats, the system will store data locally as well as have the capacity for rapid sharing regionally (e.g., to campus-based HPC centers and collaborators throughout the VOEIS community) and globally (e.g., international colleagues, TeraGrid) via UIs or programmatic APIs. This sharing will be the basis for bi-directional flows of data between storage and analysis, simulation, and eventually visualization components. This interoperability among storage and science user components of the VOEIS Data Hub will allow for rapid iterative exchange among different types of data, models, and user components. At this time we plan to implement WaterML communications for VOEIS by implementing a WaterML gateway (see section III G) to the VOEIS server. The design specifies a gateway similar to the HIS-gateway (described below) implemented through a custom DataMapper REST adapter. The WaterML gateway will respond to RESTful WaterML formatted queries within the context of a VOEIS project and will use the existing API security protocols to ensure data integrity. WaterML is currently a candidate standard in the Open Geospatial Consortium (OGC) [14] for the representation of in-situ hydrological data. WaterML 2.0 makes use of the OGC Observations and Measurement (O&M) standard [15]. The success of ecological informatics is highly dependent on the usage of common standards and the goals of VOEIS include continued support of these standards.

G. HIS Gateway

A specific goal of the VOEIS project is to integrate with the CUAHSI HIS through HydroServer. In order for VOEIS to leverage the power of the CUAHSI HydroServer and its

corresponding suite of tools, a fully functional REST interface that allows for the pulling and pushing of data objects was needed. We have constructed a HIS gateway using Ruby, Sinatra [21] and the DataMapper ORM to provide the necessary REST functionality. Associated requirements for the HIS gateway include a simple authorization system to prevent malicious access, and the ability to serve up JSON [10] and XML results from simple URL style queries that can behave like full APIs to the ODM data-store. The initial implementation of the VOEIS HIS gateway is currently available for use as a standalone JRuby [11] server application that can be deployed on any platform and configured to connect to any ODM database.

H. *Workflows and Data Provenance*

Preprocessing of data involves development of data paths through a standardized workflow framework. During quality control, errors are corrected, missing data are annotated, and metadata are created. This provides robust validation and tracking of original data which is required for comprehensive, reliable analysis later in the data workflow. More advanced workflow support with features offered by tools such as NCSA's CybeIntegrator is also currently being investigated. Research groups have well-developed data management systems and protocols. However, as typical throughout the sciences, these protocols have been largely developed to fit the specific needs of the research group. VOEIS will integrate these separate systems into a single, interactive management platform and implement data provenance (processing history) tracking. VOEIS (through its underlying implementation Yogo Framework technology) maintains data provenance by natively versioning all data stored in the system. As a result, for each VOEIS project, the data that is stored is never revised or erased. When raw data is modified (through the QA/QC process, or by some other means) the prior values are stored for provenance. Any time a change is made to any record in VOEIS a copy-on-write is performed of the pre-modified record (with the addition of a user-defined comment on the change) and is stored in a version table associated with the model. These version records are time-stamped with creation dates allowing the system to identify when any record was versioned and what version of the other records it was associated with. Therefore, as any piece of information is modified it is possible for VOEIS to ensure that for a given time the entire system could be reconstructed. Since the previous versions are read-only, the interface only allows for them to be viewed; they cannot be edited. The most recent versions of the data are the values that are used for data mining and exposed via the data retrieval APIs.

IV. CASE STUDIES

A number of projects are currently exercising VOEIS functionality and the numbers are expected to grow. This section provides a brief description of the types of projects that are or will be using VOEIS data management capabilities.

A. *The Spanish Creek Case Study*

The Spanish Creek site on the Flying D Ranch, Montana, is instrumented in support of undergraduate education at Montana State University. We are using VOEIS to instrument the site with four real-time nodes. The physical and chemical characteristics of the stream are monitored to provide data for use in Poole's class "Stream Restoration Ecology." Parameters include river stage, temperature, conductivity, dissolved oxygen, precipitation, wind speed, and incoming solar radiation. In past years, students in the class have conducted individual research projects on the creek, which have provided the foundation for each year's class to compile an integrated set of stream restoration recommendations and present them to ranch staff. Data for the VOEIS network nodes on Spanish Creek will help next year's students determine how land management and any restoration actions may be affecting the physical and biological aspects of water quality in the creek.

B. *The Tenderfoot Creek Experimental Forest Case Study*

The Tenderfoot Creek Experimental Forest (TCEF) is located in central Montana within the Lewis and Clark National Forest. As a result of collaborative hydrological and meteorological research, efforts between the Watershed Hydrology Lab at Montana State University (MSU) and the USFS Rocky Mountain Research Station, several hydrologic and meteorological monitoring stations have been set up within the forest. These include eleven streamflow gauging stations (flumes and open channel) and two eddy-covariance towers (tower at 40 m height and tripod at 3 m height). The streamflow stations collect stage, temperature, and conductivity data for each major tributary in the forest. Several sensors installed on the eddy-covariance towers assist with discerning ecosystem trends by measuring concentrations of water vapor, concentrations of carbon dioxide, and three-dimensional wind speed every tenth of a second.

Data can only be physically stored within data loggers at each station for a maximum of two to three months and is manually downloaded by Forest Service employees and MSU researchers for analysis. Access to the stations is difficult particularly in the winter months when snowmobiles must be used for travel within the forest. A remote communication system will provide a method for direct data transmission between TCEF and the VOEIS system at MSU. This will greatly reduce the potential for lost data, eliminate costly man-hours spent in the field, and will provide real-time data streams to forest managers and researchers. Such a system will provide the ability to monitor sensor activity and system power supplies as well as make watershed process predictions based on the real-time data.

C. *The Timberlake Case Study*

The Timberlake Observatory for Wetland Restoration (TOWeR) is a 440 ha former agricultural field on the coastal plain of NC that was recently abandoned, purchased by investors, and restored to a forested wetland for use as a wetland mitigation bank. This case study is a collaborative effort between Duke University, Wright State University and

Montana State University. As a result of hydrologic reconnection coupled with severe hydrologic drought, seasonal saltwater intrusion (via surface water) was documented for the first time in this site in 2007. It is anticipated that over time, these seasonal shifts will increase in both duration and salinity until ultimately TOWeR transitions to an estuarine ecosystem. Throughout this transition biogeochemical cycling will shift dramatically, but the rate and shape of this change is uncertain. Indeed, salt water intrusion and sea level rise introduce key challenges to basic understanding of coastal wetland biogeochemistry worldwide. The VOEIS system at MSU will be used to catalog and store data and metadata collected at the site. Researchers have instrumented a total of 43 permanent sampling stations throughout the site. Sampling sites are arrayed to encompass the full gradient of elevation across the site in order to capture natural variations in water levels.

V. Future Work

A challenge in ecological analysis is projecting ecological processes through time (past-to-present-to future) and across space (from field sensor locations across large landscapes). Simulation modeling provides the means for such projection. Data from sensors are used to both parameterize and to validate these models. VOEIS will be specifically designed to allow data resources to be accessed by simulation models. Collaborations with the University of Kentucky's visualization labs are currently underway to develop APIs that will facilitate high performance visualizations of these simulations. In particular, the VOEIS team is investigating the use of the Open Modeling Interface (OpenMI) [16] as a means to exchange data between operational models, thus facilitating data interchange at run time. Further, we are investigating leveraging technologies from the e-Science community (e.g., NCSA, DataOne) to avoid replication of services or the proliferation of unnecessary technologies. For example, client APIs, workflow management and provenance components.

VI. Conclusions

VOEIS provides cyber-infrastructure capabilities for managing various workflows and providing a data model that are directly aligned with the goals of other efforts currently undertaken by research and investigative groups working to promote an integrated environment for the sharing of scientific knowledge. The informatics system developed through this project is designed to manage vast amounts of legacy data as well as new data generated by the sensor networks deployed at the biological stations.

The development of the VOEIS framework enables a unique capability for PIs to manage and analyze information quickly. The informatics framework aligns itself with existing and broader efforts currently under development by the greater e-Science community.

Acknowledgments

This research is made possible by the National Science Foundation (NSF) Montana EPSCoR American Recovery and Reinvestment Act program with grant award M66012/66013. The Timberlake project research is made possible by NSF grant award 1021001.

References

[1] D.P. Ames, J.S. Horsburgh, J.L. Goodall, T. Whiteaker, D.G. Tarboton, D.R. Maidment, "Introducing the open Source CUAHSI Hydrologic Information System Desktop Application (HIS Desktop)," 18th World IMACS/MODSIM Congress, Cairns, Australia 13-17 July 2009.

[2] Data Observation Network for Earth (DataOne), http://www.dataone.org

[3] The Dojo Toolkit, http://www.dojotoolkit.org

[4] J. Futrelle, J. Gaynor, J. Plutchak, J.D. Myers, R.E. McGrath, P. Bajcsy, J. Kastner, K. Kotwani, J.S. Lee, L. Marini, R. Kooper, T. McLaren, Y. Liu, "Semantic Middleware for E-science Knowledge Spaces," Proceedings of the 7th International Workshop on Middleware for Grids, Clouds and e-Science, MGC'09, Urbana Champaign, IL., pp. 1-6, Nov. 30th – Dec 1st, 2009.

[5] J.L. Goodall, D.R. Maidment, "A spatiotemporal data model for river basin-scale hydrologic systems," International Journal of Geographical Information Science, 23:2, pp. 233-247, 2009.

[6] J.S. Horsburgh, D.G. Tarboton, D.R. Maidment, I. Zaslavsky, "A relational model for environmental and water resources data," Water Resources Res., 44, W05406, doi:10.1029/2007WR006392, 2008.

[7] C. Izurieta, G. Poole, R.Payn, I. Griffith, R. Nix, "Development and Application of a Simulation Environment (NEO) for Integrating Empirical and Computational Investigations of System-Level Complexity," unpublished.

[8] G. A. Jacobs, R.C. Heimbuch, R.R. Lamb, P. M. Llovet, S.B. Cleveland, I.R. Judson, "User Driven Evolvable Schemas," unpublished.

[9] G. A. Jacobs, R.C. Heimbuch, R.R. Lamb, P. M. Llovet, S.B. Cleveland, I.R. Judson, "The Yogo Framework: An Open Source Platform For Specific Data Management Applications," unpublished.

[10] Java Script Object Notation (JSON), http://www.json.org

[11] JRuby, http://www.jruby.org

[12] MongoDB, http://www.mongodb.org

[13] MySQL, http://www.mysql.com

[14] Open Geospatial Consortium, http://www.opengeospatial.org

[15] OGC Observations and Measurement Standard, http://www.opengeospatial.org/standards/om

[16] Open Modeling Interface, OpenMI, http://www.openmi.org

[17] Persevere, http://docs.persvr.org

[18] G. Poole, "A Data Model for Integrating Field Observations, Analytic Laboratory Results, and Data Logger Output," unpublished.

[19] PostgreSQL, http://www.postgresql.org

[20] Ruby on Rails, http://www.rubyonrails.org

[21] Sinatra Quora, http://www.quora.com

[22] SQLite, http://www.sqlite.org

[23] D.G. Tarboton, J.S. Horsburgh, D.R. Maidment, T. Whiteaker, I. Zaslavsky, M. Piasecki, J. Goodall, D. Valentine, T. Whitenack, "Development of a Community Hydrologic Information System," 18th World IMACS/MODSIM Congress, Cairns, Australia 13-17 July 2009

[24] The Unified Modeling Language (Object Management Group's UML), http://www.uml.org

[25] J. Wang, J.M. Hassett, T.A. Endreny, "An object oriented approach to the description and simulation of watershed scale hydrologic processes," Computers and Geosciences, 31, pp. 425-435, 2005.

[26] Yogo Downloads, http://github.com/yogo/yogo

[27] Zaslavsky, I., D. Valentine, and T. Whiteaker (2007), CUAHSI WaterML, Open Geospatial Consortium Discussion Paper, OGC 07-041r1, Version 0.3.0. http://portal.opengeospatial.org/files/?artifact_id=21743

Coral sensor network at Racha Island, Thailand

Mullica Jaroensutasinee[1], Krisanadej Jaroensutasinee[1], Tony Fountain[2], Michael Nekrasov[2], Sirilak Chumkiew[1], Premrudee Noonsang[1], Uthai Kuhapong[1], Scott Bainbridge[3]

[1] Center of Excellence for Ecoinformatics, School of Science, Walailak University, Nakhon Si Thammarat 80161, Thailand
[2] California Institute of Telecommunication and Information Technology, UCSD, La Jolla, CA 92093, USA
[3] Australian Institute of Marine Science, PMB 3 MC, Townsville 4810 Australia

jmullica@gmail.com, krisanadej@gmail.com, tfountain@ucsd.edu, mnekraso@ucsd.edu, sirilak.chumkiew@gmail.com, npremrudee@gmail.com, rkuthai@gmail.com, s.bainbridge@aims.gov.au

***Abstract*—Environmental observation stations are systems which allow researchers to observe rare events and to document long-term changes in ecological systems. Here we describe a system used for acquiring and sharing numerical data and imagery with ecological researchers that has been deployed at Racha Island, Phuket, Thailand. This is a new observatory that aims to provide publically accessible scientific data for researching environmental changes on coral reefs. This project is part of the Coral Reef Environmental Observational Network (CREON).**

Keywords— coral reef; sensor networks;Thailand;CREON

I. INTRODUCTION

Coral reefs are the most complex, species rich, and productive marine ecosystem [1-3]. The benefits of coral reefs are crucial to tourism, fisheries, shoreline protection, medicines and they serve as environmental indicators making them a priority for conservation and a major concern for sustainable development [e.g. 3-9]. Elevated ocean temperatures due to global warming, changes in salinity, intense solar radiation, low wind, exposure to air at low tides or low sea level, sedimentation or chemical pollutants, can cause major stress to coral and lead to coral bleaching events [10-17].

The link between physical conditions and the biological responses that lead to coral bleaching [13] allows for the prediction of when corals may bleach based on measurements of the in-situ physical parameters. Monitoring of these parameters therefore becomes an important part of understanding and responding to coral bleaching events.

The term 'Sensor Network' refers to an array of interconnected (normally using wireless technologies) small sensors that stream real time data back to a central point. The communication with the sensor is typically bi-directional allowing for 'smart' adaptive sensing, event detection and on-node information processing. Sensor networks are powerful tools for environmental monitoring including environmental data collection, pollution monitoring, disaster prevention, tsunami and seaquake warning [22, 23]. They allow for the monitoring and detection of phenomena more accurately and rapidly in a variety of geographical areas. Recently, applying sensor networks in underwater environments has received growing interest [24-29]. To improve the understanding of coral reef ecosystems, it is essential that studies are conducted over a wide range of temporal and spatial scales.

Cameras have been extensively used in ecology including observations of the nocturnal behavior of coral reef fishes [30] and the study of large cryptic animals [31-32]. Such applications take advantage of the camera's ability to provide unobtrusive observations over long time periods in inaccessible locations. Most camera deployments are only for short periods limiting the number of images that are captured and so the number and type of events recorded. A permanently installed network-connected web camera can capture a constant set of images and data indefinitely ensuring that even rare events are sampled. Similarly, networked sensors measuring physical characteristics of ecosystems, such as temperature, conductivity and pressure, can also provide high-resolution records over long time periods. Integrated sensor suites for capturing numeric and image data can generate high data rates (standard definition video has a data rate over 3.5 Mbit/s, compressed HD video is over 25Mbit/s or 5 GB per hour to store). These high data rates and the heterogeneity of the data types demand new approaches to networking, data management, visualization, and analysis [33].

Access to near real-time data during bleaching events, using sensor networks, is essential in advancing our understanding. Early warning of local conditions likely to cause coral bleaching could enhance regional alerts to assist: (1) science in documenting and researching the phenomena, (2) public relations in keeping reef-based commercial operators, politicians and the general public informed and (3) coral reef managers in ameliorating local-scale human impacts that might exacerbate coral bleaching. In this paper, we describe a coral reef sensor network at Racha Island, Phuket, Thailand.

II. MATERIALS AND METHODS

A. Study Site

This study was undertaken at Racha Yai Islands, Phuket province, Thailand (Latitude 7.60488 °N, Longitude 98.37660 °E) (Fig. 1, Google Earth). Coral reefs in this area are typically shallow (1-15 m depth) fringing reefs.

Figure 1. Racha Island, Thailand

The Racha Island site is a logistically challenging environment for both researchers and instruments, characterized by large but shallow bays, storms, and occasional power and internet outages. The climate is tropical with mean monthly temperatures that range between 25-30 °C. Note that large scale bleaching was observed at this site in 2009-2010 with some of the HOBO loggers recording water temperatures of up to 33 °C.

B. Collaboration

This project is part of the Coral Reef Environmental Observatory Network (CREON) [34], a group of international institutions made up of scientists and engineers whose goal is to develop tools for coral reef research. Building on CREON, this project is a collaboration between a diverse team of ecologists, computer scientists, and engineers from the California Institute of Telecommunications and Information Technology at the University of California San Diego (CalIT2 UCSD, www.calit2.net), the Australia Institute of Marine Science (AIMS, www.aims.gov.au) and the Center of Excellence of Ecoinformatics, NECTEC-Walailak University. This deployment builds on the experiences of CREON members in establishing coral reef observatories that share and interchange data from multiple sites around the Pacific Rim. It is envisioned to be a living laboratory for long-term studies of marine ecology and as a test-bed for evolving technologies for environmental and biological sensing, communications, and analysis.

C. Instruments and Infrastructure

The following description of the current deployment is organized into three areas: field deployment, cyber-infrastructure, and visualization.

1) Field Deployment

At the field site, there are a variety of aquatic and terrestrial sensors that provide a comprehensive view of the environment for coral reef ecology. All of these instruments are commercially available and widely used by the marine sciences community (Table 1).

TABLE I. DEPLOYED SENSORS IN REAL-TIME SYSTEM

Sensor	Sampling Interval	Measurements	Networked
Weather Station	1 min	Temperature, Rain, Wind, Humidity, Bar. Pressure, Solar Radiation	Yes
CTD	5 min	Conductivity, Temperature, Depth	Yes
HOBO	10 min	Temperature, Light	No
EcoCam	Continuous	Video	Yes

On June 2007, HOBO Pendant temperature and light data loggers (UA-002-64) were deployed to measure water temperature and light intensity with a 10 min sampling frequency. These sensors are not networked and require a diver to collect data every three months.

In November 2009, a Davis Vantage Pro II Plus weather station for measuring air temperature, rainfall, wind, barometric pressure, UV index and solar radiation was installed on shore with a 1 min sampling frequency.

On February 2010, four EcoCams capable of real time video capture were deployed, one underwater on the reef and three on land. The cameras provide researchers and students with a real time view of the reef and surrounding environment.

In October 2010, a SeaBird SBE37 conductivity (salinity), temperature and depth (via pressure) sensor package, commonly referred to as a CTD, was deployed on the fringing reef in approximately 10 m water depth with five minute sampling frequency. The deployment uses inductive coupling technology to send the data back to the station on the shore. A 350 m plastic coated steel cable (mooring wire) runs from shore to the CTD, secured at 10 m intervals by three kg cinder bricks. The CTD is connected to the mooring cable via an inductive modem connection. In the future, additional sensors can be attached to this cable to provide additional measurements without needing to change the cabled network infrastructure. This system provides a scalable and robust foundation for communication between sensors and the on-shore data processing computer.

2) Cyber-infrastructure

The weather station, CTD, and EcoCams stream observations in real time to a Data Center located at Walailak University (WU) and mirrored to UCSD and Nakhon Si Thammarat Rajabhat University (NSTRU). The system includes cyber-infrastructure for real-time streaming data acquisition, scalable event stream processing, and data publication services. Scientists at WU, UCSD, AIMS and other remote locations access the data and event streams via a suite of client applications for visualization, modeling, and analysis. The system is engineered to be scalable, robust, extensible, and secure. It is built using state-of-the-art open-source software tools.

The acquisition and transfer of data is accomplished using DataTurbine, a real-time streaming data engine [14]. It is an open-source middleware product supported by NSF, NASA, and private industry managed by the NSF-sponsored Open Source DataTurbine Initiative at CalIT2 (www.dataturbine.org). The DataTurbine middleware satisfies

a core set of infrastructure requirements that are common in environmental observing systems, including reliable data transport, a framework for integrating heterogeneous instruments, and a comprehensive suite of services for data management, routing, synchronization, monitoring, and visualization [35,36]. From the perspective of distributed systems, the DataTurbine middleware is a "black box" to which applications and devices send and receive data. DataTurbine handles all data management operations between data sources and sinks, including reliable transport, routing, scheduling, and security. DataTurbine accomplishes this through the innovative use of flexible network bus objects combined with memory and file-based ring buffers. Network bus objects perform data stream multiplexing and routing. Ring buffers provide tunable persistent storage at key network nodes to facilitate reliable data transport.

In addition to DataTurbine, a secondary system for storing video data is used. In conjunction with the cameras, the submersible underwater monitor system (CR110-7) and Recorder DVR (FK-RJ2604) provide a high frequency feed for live observation, with periodic archiving of images for retrospective analyses. Live online feeds provide updated images every 5 s, which is a compromise between researcher needs and camera capabilities. Archive images are typically taken every three hours. Files are transferred real-time online into an FK-RJ2604 DVR device.

Data from the DataTurbine server are extracted and uploaded to a database on a regular basis (daily) and this forms the long term data store for the project. Data from the logging HOBO sensors are manually uploaded to this database after every download (three monthly) to produce a final integrated data set. The time codes stored in the database can be manually matched to the video to link the visual data to the physical data; work is underway to automate this process so that for any set of physical measurements the video can be automatically viewed.

3) Visualization

This site uses a variety of techniques to visualize and share data. Our primary objective in creating this site was to make information freely and easily accessible both to ecological researchers and school students. All the research work is documented and photographed, and activities, as well as results of research are published to http://www.twibl.org/virtualsites/ for use by schools. The video streams are accessible through a website by researchers and students.

All data are also accessible through the DataTurbine as well as a number of client applications that interface with DataTurbine that can be run remotely. Some of these operate on real-time data streams; some operate on the archived data. These include the DataTurbine Real-time Data Viewer (RDV), a utility for creating embedded web page graphs, a MATLAB interface, and a Google Earth plug-in.

Through DataTurbine, users can see temporally synchronized streams of both video and numeric data allowing researchers to match environmental variables on air and water with pictures, providing context. There are also plans to utilize the DataTurbine services to build a web site for the Racha Island observatory to make it easier for the public to interact with the data in real time.

III. RESULTS

Since becoming operational the system has provided scientists with significant new insights into the coral reef ecosystem at Racha Island. The data collected by the system includes key physical parameters such as in water and above water temperature, water salinity and meteorological conditions (Fig. 2) as well as above water and in-water video images (Fig. 3).

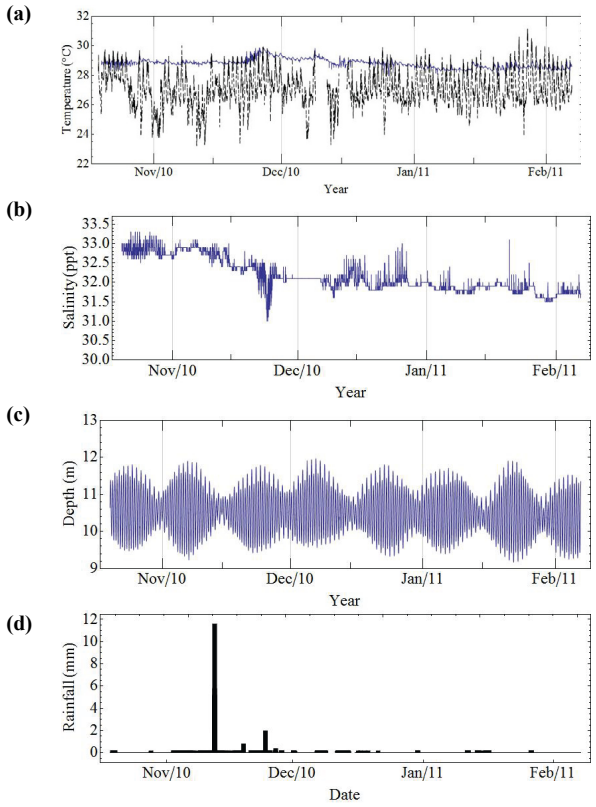


Figure 2. Physical conditions at Racha Island, Thailand from 19 October 2010-6 February 2011. (a) water temperature at 10 m (°C) (solid line), air temperature (°C) (dashed line), (b) salinity (PSU), (c) depth (m) and (d) rainfall (mm).

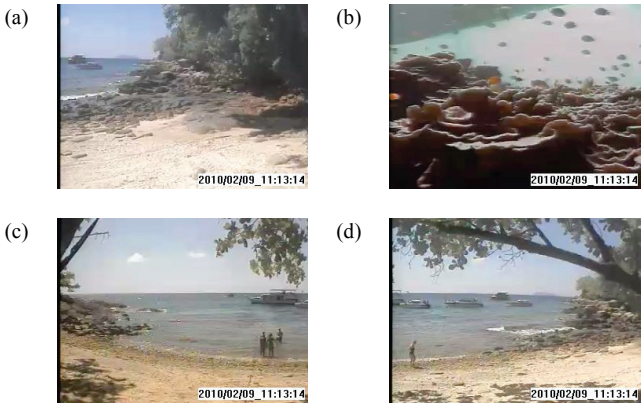


Figure 3. EcoCams stream observations at Racha Islands, Thailand. (a, c, d) Racha Island beach and (b) coral reef site.

IV. DISCUSSION

The system has been operational since coming on line in 19th October 2010. The Data Center services have been very stable. The only interruptions were for scheduled system maintenance and power outage. The field data acquisition system has also been stable, although as the system is new occasional user interaction has been required due to some initial growing pains that come with deploying a new system. The system has been robust to occasional power and network outages, even during through several very heavy storms in early November 2010.

Racha Island had extensive coral bleaching in 2010 when the HOBO temperature loggers recorded temperatures over 33 °C (personal observation, authors MJ and KJ). The data from 2011 shows much lower temperatures (maximum of 30.6 °C.) with a result that no bleaching has been observed this year for this site.

In one application of the imagery, researchers plan to sample images every 10 min to count the number of coral reef fish, and to determine the results of the interactions between coral and coral reef fish (feeding rate, aggressive behavior, tourist impact and etc). The camera systems as described here provide new capabilities for ecologists studying a wide range of phenomena. They facilitate high-frequency monitoring over long time spans which allowed them to capture infrequent events that would otherwise have gone unobserved. The infrequent events would have been impossible using a human observer both due to the cost of paying the observer and because the presence of a human so close to the observing site would have altered the dynamics of animal interactions.

Although the camera systems presented here have proved useful for ecological research, there remain many additional challenges and opportunities. Some specific challenges that need to be overcome include variable lighting intensity and angle, plasticity in the size, configuration and orientation of features of interest, the wide diversity of possible features of interest and even mundane problems such as environmental fouling as algae collects on the lens of the remote camera. However, if such challenges can be surmounted it opens additional opportunities for automated or semi-automated data collection using web cameras.

The infrastructure allows for adaptive sampling in that sampling rates can be altered based on the data being collected. While this was possible the lack of standard interfaces to each of the instruments and the need to write considerable code to automate this meant that the adaptive sampling functionality was achieved by manually re-programming the instruments. This is cumbersome and while it can be done remotely it is time consuming and only practical in a small scale deployment such as Racha Island. This is an area that is still unresolved and where common instrument interfaces and programming protocols would help.

The work being done fits within the larger CREON group and within this group solutions for higher level data management are being investigated. These include a single cloud-based data store for data from each of the CREON sites, metadata for all sensors in ISO-19115 format [37] and web based access and analysis tools. While this work is on-going the outcome will be a single system that will allow for comparisons to be done across the CREON sites and an ability to better understand the factors impacting coral reefs including responses such as coral bleaching.

V. CONCLUSION

Understanding the processes that impact reefs, such as temperature, requires high quality data at a range of spatial scales on a regular basis. Autonomous smart sensor based systems provide one way to obtain these data from the scale of oceans to the scale of individual corals. The development of a suite of technologies to deliver a robust, simple but effective technology platform to support sensor webs has become a high priority for a number of marine and environmental agencies. This project looks to take this goal forward for coral reefs using a number of technologies and a number of partners. Some of the technical obstacles are similar for any marine based monitoring system and mainly revolve around fouling, powering equipment and the general problems of maintaining equipment in a remote and hostile (at least to electronics) environment.

There are, however, a number of new challenges that need to be addressed. This include being able to store and deal with the large amounts of data that the system will generate (which may include video feeds), the integration of the data into modeling and visualization systems and the ability to manage and maintain a system that is inherently more complex than the simple passive systems deployed currently. We hope our efforts will create a valuable technology knowledge base for the further deployment of reef monitoring systems in remote environment.

The cost of sensor networks needs to be considered. While the individual elements are not expensive, systems of this type will only realize their true potential if they are replicated in large numbers. We are as far as possible employing off the shelf or simple to fabricate hardware and software solutions.

The work done opens opportunities for the development of lower cost systems that, through the use of consumer grade electronics (such as smart-phones), advances in the development of more cost effective sensors and through the work being done on open-source data management and visualization software. One goal is to develop systems that can be deployed simply and at a reasonable cost to dramatically increase the number of units that can be deployed and correspondingly our understanding of the processes that sustain and threaten coral reef systems.

ACKNOWLEDGMENTS

We thank Damien Eggeling and Geoff Page from AIMS for the CTD installation, and Banraya Resort and Spa staff for providing the research facilities at Racha Island. This work was supported in part by the United States National Science Foundation, AIMS, and the Center of Excellence for Ecoinformatics, the Institute of Research and Development, NECTEC/Walailak University.

REFERENCES

[1] J. Stafford-Deitsch, "Reefs: a safari through the coral world". Sierra Club Books, San Francisco, 1993.

[2] K.P. Seben, "Biodiversity of coral reefs: what are we losing and why?, *Am. Zool*., 1994, 34, 115-133.

[3] D. Bryant, L. Burke, J. McManus, and M. Spalding, "Reefs at risk: a map-based indicator of threats to the world's coral reefs. World Resources Institute, Washington D.C. 1998.

[4] R.L. Hayes, and T.J. Goreau, "Tropical coral reef ecosystems as a harbinger of global warming", *World Resource Rev*., 1991, 3, 306-322.

[5] S.C. Jameson, J.W. McManus, and M.D. Spalding, "State of the reefs: regional and global perspectives", U.S. Department of State, Washington DC. 1995.

[6] R.H. Richmon, "Coral reefs: present problems and future concerns resulting from anthropogenic disturbance", *Am. Zool*., 1993, 33, 524-536.

[7] T.J. Done, J.C. Ogden, and W.J. Wiebe, "Biodiversity and ecosystem function of coral reefs", In H.A. Mooney, J.H. Cushman, E. Medina, O.E. Sala, and E.D. Schulze (eds), Functional roles of biodiversity: a global perspective. Wiley, Chichester, UK. 1996, pp. 393-429.

[8] W. Fenical, "Marine biodiversity and the medicine cabinet: the status of new drugs from marine organism. *Oceanography*, 1996, 9, 23-27.

[9] J.K. Reaser, R. Pomerance and P.O. Thomas, "Coral bleaching and global climate change: scientific findings and policy recommendations", *Conserv. Biol*., 2000, 14, 1500-1511.

[10] J. Oliver, "Recurrent seasonal bleaching and mortality of corals on the Great Barrier Reef", *Proc. 5th Int. Coral Reef Symp*., 1985, 4, 20-206.

[11] M.W. Gleason and G.M. Wellington, "Ultraviolet radiation and coral bleaching", *Nature*, 1993, 365, 836-838.

[12] P. Glynn. "Coral reef bleaching: facts, hypotheses, and implications", *Global Change Biol*., 1996, 2, 495-509.

[13] G.D. Dennis, and E.I. Wicklund, "The relationship between environmental factors and coral bleaching at Lee Stocking Island, Bahamas in 1990", In Ginsburg R.N. (compiler). Proc. Colloq on Global Aspects of Coral Reefs: health, hazards and history, 1993. Rosentiel School of Marine and Atmospheric Science, University of Miami, pp. 167-173.

[14] J.H. Drollet, M. Faucon, S. Maritorena, and P.M.V. Martin, "A survey of environmental physic-chemical parameters during a minor coral mass bleaching event in Tahiti in 1993". *Aus. J. Mar. Freshw. Res*., 1994, 45, 1149-1156.

[15] M.D. McField, "Coral response during and after mass bleaching in Balize", *Bull. Mar. Sci.* 1999, 64, 155-172.

[16] S.D. Donner, W.J. Skirving, C.M. Little, M. Oppenheimer, and O. Hoegh-Guldberg, "Global assessment of coral bleaching and required rates of adaptation under climate change". *Glob. Change Biol.*, 2005, 11, 2251-2265.

[17] M.J.C. Crabbe, "Global warming and coral reefs: Modelling the effect of temperature on *Acropora palmate* colony growth", *Comput. Biol. Chem*., 2007, 31, 294-297.

[18] J.M. Lough, "Climate variability and change on the Great Barrier Reef", In E. Wolanski (ed) Oceanographic processes of coral reefs: physical and biological links in the Great Barrier Reef, CRC Press, Boca Raton, FL, 2001.

[19] S.L. Coles, P.L. Jokiel, and C.R. Lewis, "Thermal tolerance in tropical versus subtropical Pacific reef corals", *Pac. Sci*., 1976, 30, 159-166.

[20] J. Marcus, and A. Thorhaug, "Pacific versus Atlantic responses of the subtropical hermatypic coral *Porites* spp. to temperature and salinity effects, *Proc. 4th Int. Coral Reef Symp*., 1981, 2, 15-20.

[21] P.W. Glynn, J. Cortes, H.M. Gusman, and R.H. Richmon, "El Niño (1982-83) associated coral mortality and realtionsip to sea surface temperature deviations in the tropical eastern Pacific", *Proc. 6th Int. Coral Reef Symp*., 1898, 3, 237-243.

[22] P. Xie, Z. Zhau, Z. Peng, J.-H. Cui, and Z. Shi, "SDRT: a reliable data transport protocol for underwater sensor networks", *Ad Hoc Networks*, 2010, 8, 708-722.

[23] D. Pompili, T. Melodia, and I.F. Akyildiz, "Three-dimensional and two-dimensional deployment analysis for underwater acoustic sensor networks", *Ad Hoc Networks*, 2009, 7, 778-790.

[24] I.F. Akyildiz, D. Pompili, and T. Melodia, "Challenges for efficient communication in underwater acoustic sensor networks", *ACM SIGBED Rev.*, 2004, 1, 1.

[25] P. Xie, J.-H. Cui, and L. Lao, "VBF: vector-based forwarding protocol for underwater sensor networks", Proceedings of IFIP Networking' 06, Coimbra, Portugal, May 2006.

[26] M. Chitre, S. Shahabudeen, and M. Stojanovic, "Underwater acoustic communication and networks: recent advances and future challenges", *Mar. Technol. Soc. J.*, 2008, 1, 103-116.

[27] J.-H. Cui, J. Kong, M. Gerla, and S. Zhou, "Challenges: building scalable applications", *IEEE Network Special Issue on Wireless Sensor Networking*, 2006, 20, 12-18.

[28] J. Heidemann, Y. Li, A. Syed, J. Wills, and W. Ye, "Research challenges and applications for underwater sensor networking", *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, 2006, 12-18.

[29] L. Liu, S. Zhou, and J.-H. Cui, "Prospects and problems of wireless communication for underwater sensor networks", *Wireless Communication and Mobile Computing*, 2008, 8, 977-994.

[30] S.J. Holbrook, and R.J. Schmitt, "Settlement patterns and process in a coral reef damselfish: in situ nocturnal observations using infrared video", *Proc. 8th Int. Coral Reef Symp*., 1997, 2, 1143–1148.

[31] K.U. Karanth, and J.D. Nichols, "Estimation of Tiger densities in India using photographic captures and recaptures", *Ecology,* 1998, 79, 2852-2862.

[32] L. Silveira, A.T.A. Jácomo, and J.A.F. Diniz-Filho, "Camera trap, line transect census and track surveys: a comparative evaluation", *Biol. Conserv.*, 2003, 114, 351-355.

[33] E. Strandell, S. Tilak, H.-M. Chou, Y.-T. Wang, F.-P Lin, P. Arzberger, T. Fountain, T.-Y. Fan, R.-Q. Jan, and K.-T. Shao, "Data Management at Kenting's Underwater Ecological Observatory", *3rd Int. Conf. Intelligent Sensors, Sensor Networks and Information, Melbourne*, 2007.

[34] CREON: The Coral Reef Environmental Observatory Network, http://www.coralreefeon.org/

[35] T. Fountain, S. Tilak, P. Hubbard, P. Shin, and L. Freudinger, "The Open Source DataTurbine Initiative: Streaming data middleware for environmental observing systems," *33rd Int. Symp. on Remote Sensing of Environmen*t, Stresa, Italy, 4-8 May 2009. (http://www.dataturbine.org/biblio)

[36] S. Tilak, P. Hubbard, M. Miller, and T. Fountain, "The Ring Buffer Network Bus (RBNB) DataTurbine Streaming Data Middleware for Environmental Observing Systems," *Proc. 3rd IEEE Int. Conf. E-Science and Grid Computing (e-Science)*, 2007.

[37] International Standards Organization, "ISO 19115:2003 Geographic information – Metadata," ISO 19115:2003(E), 2003.

The Process of Digitising Natural History Collection Specimens at Digitarium

Juha Lehtonen, Susanne Heiska, Mika Pajari, Riitta Tegelberg & Hannu Saarenmaa
Digitarium - Digitisation Centre of the Finnish Museum of Natural History and the University of Eastern Finland
Faculty of Science and Forestry, Joensuu Science Park
Länsikatu 15 (P.O. Box 111), FIN-80101 Joensuu
www.digitarium.fi hannu.saarenmaa@uef.fi

Abstract— **Digitarium is a joint initiative of the Finnish Museum of Natural History and the University of Eastern Finland. It was established in 2010 as a dedicated shop for large scale digitisation of collections. The paper gives an overview of the steps of digitisation process, including tagging, imaging, data entry, georeferencing, filtering, validation, publishing, and archiving. A functional model is presented. The work at Digitarium is independent of any collection management software. Instead, the digitisation process is managed through XML-documents and versioning. All specimens are imaged and distance workers take care of the digitisation from the images. Data and images are published through Morphbank and GBIF.**

Keywords— digitisation; imaging; natural history collections; XML.

I. INTRODUCTION

Digitisation of specimens in natural history collections is a huge challenge (cf. [1]) . A total of 2-3 billion specimens has been estimated to exist in natural history museums worldwide, and less than 5% of them have been catalogued digitally until now. A much smaller proportion has been imaged. In Finland, the six largest public museums contain an estimated 22 million specimens, out of which 12% has been digitally catalogued, i.e. minimally digitised. In addition, private collections contain up to 8 million specimens.

In the national digitization strategy of natural history collections [2] it has been estimated that the required effort to digitize most of these holdings is 750-1000 person years. This estimate is based on the rate of 100 samples per day for each worker. Such rate has been reported in well-organised digitization projects [1, 2] with easy materials. However, the current efficiency in most digitisation projects still seems to be around 20-30% of this optimum. These rates need to be improved, if digitisation at all is going reach its goals. The improvement should come from moving from hand-crafting to industrial-scale assembly lines and workflows.

As a response to this challenge, Digitarium, the Digitisation Centre of the Finnish Museum of Natural History and the University of Eastern Finland was established in 2010. Digitarium implements the national digitisation strategy for natural history collections [2], and aims at speeding up of digitisation through an efficient production line and knowledge management.

This document outlines the process of digitisation as it is being implemented at Digitarium, and the related ICT support. Not all the steps are fully in place at this writing, but the production line works, and is being streamlined and tuned up.

Special features of the process at Digitarium are imaging of all material, distributed workflow that can employ distance workers, and XML (Extensible Markup Language) based data management. The process is for the first time being described here.

II. PROCESS STEPS

The below steps normally occur in sequence in this order, and are driven by the JJC tool, which is described below. They are illustrated in the functional model of Fig 1.

A. Receiving

Digitarium does not manage its own collections, but is a shop for digitising materials from "customers", i.e., museums and other institutions located elsewhere. Therefore, delivery of material is the first step. An agreement is made with the sending institute of the material that will be received, and in what detail and timeframe it will be processed. After receiving, material is subjected to deep-freezing to eliminate any pest organisms. A metadata entry is made about the received material and agreements.

B. Tagging

Each sample will be tagged with a globally unique identifier in the form of an HTTP URI, for example http://id.luomus.fi/GA.105. This namespace is managed by the Finnish Museum of Natural History. The URI is resolvable to the specimen details. The last part of the URI will also be written in a 3-dimensional barcode. The tag will be glued to the paper sheet or pinned in the needle of an insect sample. The labels from insect specimens are removed and placed temporarily on a sheet of cardboard for imaging.

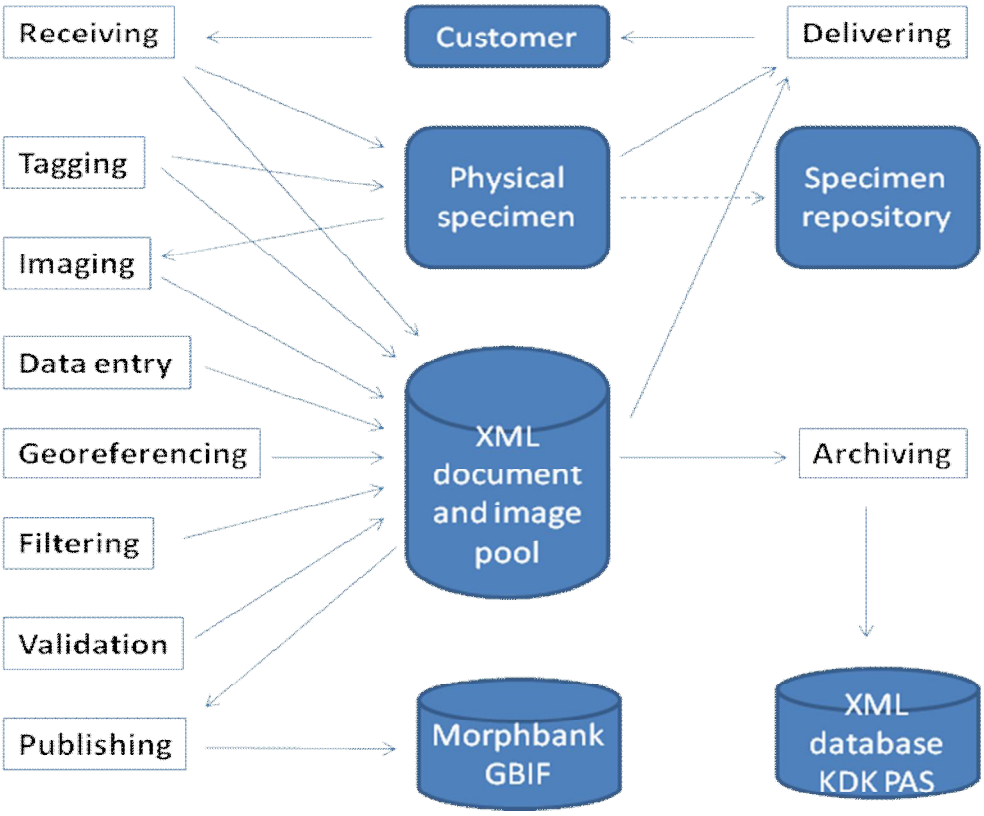


Figure 1. Functional model of the digitisation process.

Figure 2. An example of the result of imaging the lower half of a plant sheet, which is the source of the data in Fig. 3.

C. *Imaging*

Several pictures are made of the sample with a high-end digital camera. The cameras have resolution of 24 megapixels and produce TIFF images of 75 MB. A plant sheet is imaged in two pieces (Fig 2.), which gives a resolution of 450 dpi over the entire sheet. The two pieces are later joined programmatically. In the case of insect samples, the specimen and the labels are imaged separately. Scanners are not employed. Details of the imaging event and results are stored in an XML document automatically.

```
<?xml version="1.0" encoding="UTF-8"?>
<dwr:DarwinRecordSet
xmlns:xsi="http://www.w3.org/2001/XMLSchema"
xsi:schemaLocation="http://rs.tdwg.org/dwc/dwcrecord/
http://rs.tdwg.org/dwc/xsd/tdwg_dwc_classes.xsd"
xmlns:dcterms="http://purl.org/dc/terms/"
xmlns:dwc="http://rs.tdwg.org/dwc/terms/"
xmlns:dwr="http://rs.tdwg.org/dwc/dwcrecord/">
<dwc:Occurrence>
    <dwc:occurrenceID>http://id.luomus.fi/EA9.00C</dwc:occurrenceID>
    <dwc:associatedMedia>./EA9.00C/Image001.tif;
./EA9.00C/Image002.tif; ./EA9.00C/Preview001.jpg;
./EA9.00C/Preview002.jpg</dwc:associatedMedia>
    <dwc:recordedBy>Laine Matti</dwc:recordedBy>
    <dwc:preparations>dry</dwc:preparations>
    <dwc:individualCount>1</dwc:individualCount>
    <dwc:disposition>in collection</dwc:disposition>
    <dcterms:type>PhysicalObject</dcterms:type>
    <dcterms:modified>2011-05-04; 2011-03-31</dcterms:modified>
    <dcterms:creator>Pennanen, Marja (d); Lemmetyinen, Juha
(i)</dcterms:creator>
    <dcterms:contributor>Digitarium</dcterms:contributor>
    <dcterms:language>FI</dcterms:language>
    <dwc:basisOfRecord>PreservedSpecimen</dwc:basisOfRecord>
</dwc:Occurrence>
<dwc:Event>
    <dwc:eventID>http://id.luomus.fi/EA9.00C</dwc:eventID>
    <dwc:fieldNumber>G333</dwc:fieldNumber>
    <dwc:eventDate>1986-8-31</dwc:eventDate>
    <dwc:habitat>harjun S-rinteen puoliväli</dwc:habitat>
    <dwc:year>1986</dwc:year>
    <dwc:month>8</dwc:month>
    <dwc:day>31</dwc:day>
    <dwc:occurrenceID>http://id.luomus.fi/EA9.00C</dwc:occurrenceID>
</dwc:Event>
<dwc:Identification>

<dwc:identificationID>http://id.luomus.fi/EA9.00C</dwc:identificatio
nID>
    <dwc:identifiedBy>Laine Matti</dwc:identifiedBy>
    <dwc:occurrenceID>http://id.luomus.fi/EA9.00C</dwc:occurrenceID>
</dwc:Identification>
<dcterms:Location>
    <dwc:locationID>http://id.luomus.fi/EA9.00C</dwc:locationID>
    <dwc:continent>Europe</dwc:continent>
    <dwc:countryCode>FI</dwc:countryCode>
    <dwc:stateProvince>V;Ab</dwc:stateProvince>
    <dwc:municipality>Lohjan kunta</dwc:municipality>
    <dwc:locality>Perttilä; Unionin huoltoaseman E-puolella kantatie
53:n vierellä</dwc:locality>
    <dwc:verbatimCoordinates>668  :33</dwc:verbatimCoordinates>
    <dwc:verbatimLatitude>668</dwc:verbatimLatitude>
    <dwc:verbatimlongitude>333</dwc:verbatimlongitude>
    <dwc:verbatimCoordinateSystem>YKJ</dwc:verbatimCoordinateSystem>
    <dwc:locationRemarks>"Kantatie 53" is changed in the year 1996
to "valtatie 25".</dwc:locationRemarks>
    <dwc:occurrenceID>http://id.luomus.fi/EA9.00C</dwc:occurrenceID>
</dcterms:Location>
<dwc:Taxon>
    <dwc:taxonID>http://id.luomus.fi/EA9.00C</dwc:taxonID>
    <dwc:taxonRank>variety</dwc:taxonRank>
    <dwc:scientificName>Diphasiastrum complanatum ssp.  ×
zeilleri</dwc:scientificName>
    <dwc:scientificNameAuthorship>(L.) J. Holub;  (Roy)
Kukk.</dwc:scientificNameAuthorship>
    <dwc:genus>Diphasiastrum</dwc:genus>
    <dwc:specificEpithet>complanatum</dwc:specificEpithet>
    <dwc:infraspecificEpithet>zeilleri</dwc:infraspecificEpithet>
    <dwc:taxonRemarks>ssp. means subspecies but zeilleri is a
hybrid. Author names are mispelled.</dwc:taxonRemarks>
    <dwc:occurrenceID>http://id.luomus.fi/EA9.00C</dwc:occurrenceID>
</dwc:Taxon>
</dwr:DarwinRecordSet>

Image001.tif
Image002.tif
Preview001.jpg
Preview002.jpg
```

Figure 3. Example of a sample in XML-document after the data entry phase, describing the plant sheet in Fig 2.

D. *Data entry*

The data from labels is entered manually from images using the JJC tool. The data are stored as written, including any misspellings, abbreviations, etc. into the "Verbatim" fields of the Darwin Core data exchange standard, see http://rs.tdwg.org/dwc/. A new version of the XML-document is generated, and the old version from the previous step is kept. Here, like in all other steps of the process a separate document version is retained.

E. *Georeferencing*

Most specimens do not come with geographic coordinates, and candidates for these will be found automatically using web services such as GEOLocate and those of the Finnish National Survey. This results normally in several choices, which are ranked and stored in the XML file in the Darwin Core field georeferenceRemarks. In case grid coordinates have been given in the sample using the Finnish national system (called "YKJ"), these are automatically converted into geographic coordinates already in the previous phase on data entry, and no further candidates are searched.

F. *Filtering*

Certain details of datasets sometimes need to be filtered out before publishing, because of reasons such as endangered species or the customer requiring an embargo of the material. Such filtering is done automatically based on species name or an agreement stored in the metadata of the dataset. Detailed instructions, but still in draft form, exist for this step [3]. Two versions of the XML file are retained: filtered and unfiltered.

G. *Validation*

After all the preparation above, some steps of which being automatic, a final check is made by an experienced staff member. Any errors in data entry are corrected. Out of the georeferenced location candidates one is chosen, or a new manual search is made, and data is stored in the decimalLatitude, decimalLongitude, and precision fields of Darwin Core. The result of any filtering is checked and masked versions of the images will manually be created.

H. *Publishing*

The data from the latest XML document version and images will be imported to Digitarium's Morphbank database instance and Digitarium's GBIF IPT service. From there they will be published, as agreed with the customer, or if publication has not been agreed, retained for Digitarium's internal use.

I. *Delivery*

The data will be delivered to the customer in the agreed format. The samples will be sent back, unless it has been agreed to keep and curate them at Digitarium's repository.

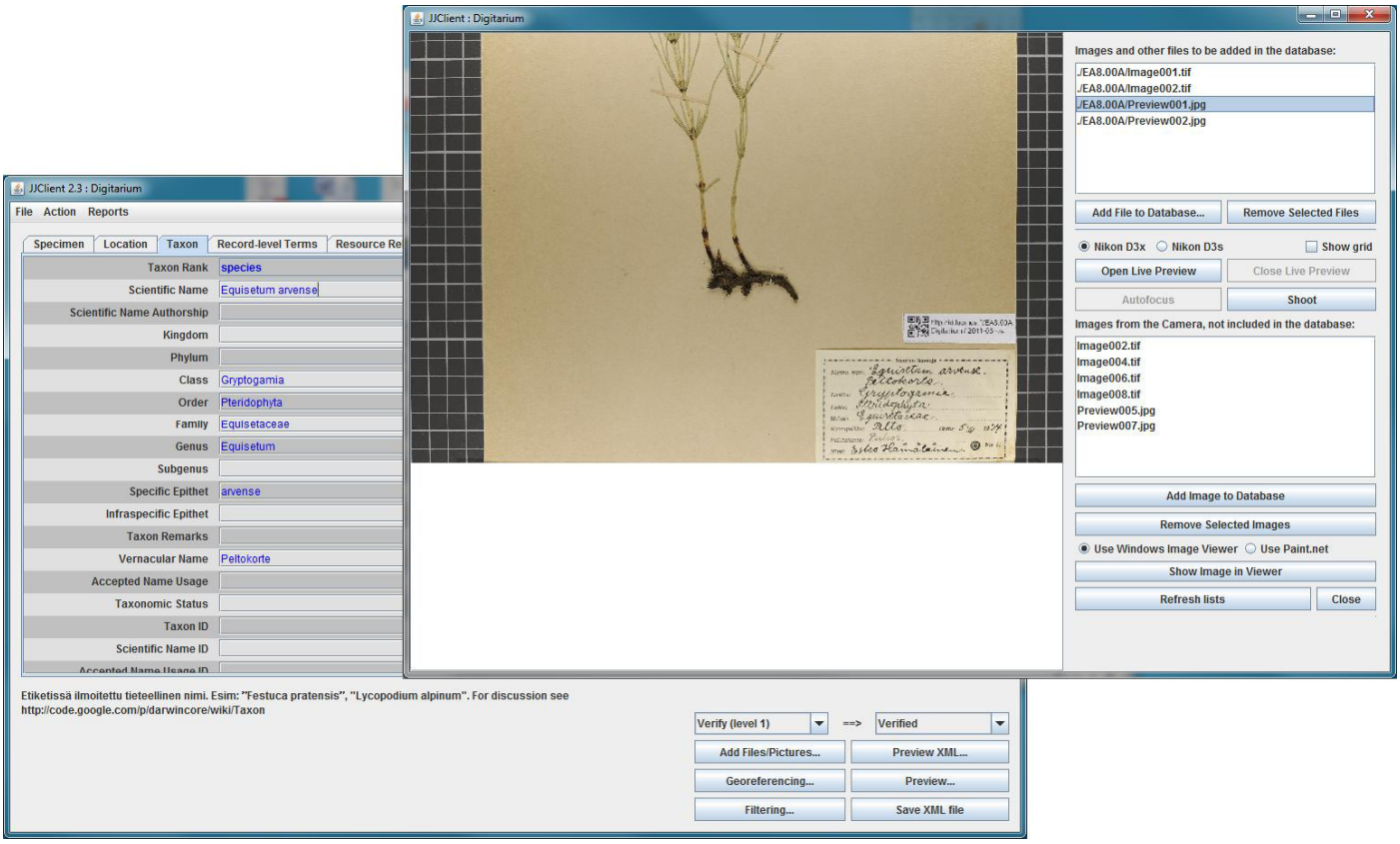


Figure 4. Snapshot of JJC showing some workflow functions.

J. *Archiving*

All the XML documents and images will be retained indefinitely on Digitarium's Metacat service and eventually at the long-term archival service of the National Digital Library (KDK PAS).

III. THE ICT SYSTEM AND WORKFLOW

The above steps are being supported by an ICT system that implements the workflow. Its main components are described below.

A. *XML and Darwin Core*

All original data is being stored in XML documents. They contain only terms from the Darwin Core and Dublin Core (http://dublincore.org/documents/dces/) standards. A detailed guide for their application at Digitarium has been written [4]. An example is given in Fig. 3. Typically a new version of the XML document is generated at each step of the process. For the time being, the all the documents and images are managed just on the file system. Metadata describing datasets (i.e. groups of Darwin Core XML documents, and orders by customers) will also be stored in XML files using the EML (Ecological Metadata Language) standard.

B. *Digitisation workbench*

A dedicated tool has been written for digitisation and automation of workflow. This tool, called JJC, has been written in Java at Digitarium by the first author. The tool runs in Windows and provides for data entry into the XML documents, and driving of Nikon cameras for imaging. It can retrieve and write the XML documents pertinent to each step in the workflow. See Fig. 4.

C. *Morphbank*

This service is available at http://morphbank.digitarium.fi/ and it is part of the global and Nordic collaboration. Morphbank is a database tool designed in particular for natural history specimens, the locations where they have been collected, images of them and their parts, image views, taxonomy, and annotations [5]. Morphbank is a publishing platform in the sense that after publishing date, the objects in principle cannot be removed from the service anymore and all objects have stable short URIs that can be reused elsewhere.

D. *XML database*

In order to keep track of all the XML documents and their versions, a document repository needs to be used. Ideally, it will support search within the XML documents. Metacat, the XML database tool from the LTER network is has been used elsewhere for long term archival and search of material and

related metadata [6]. Meta cat is not yet in production at Digitarium, but is being tested. The National Digital Library of Finland (KDK) is building a long term archival systemn (PAS), which also will be used when it becomes operational in 2016.

IV. CONCLUSION

The process and tools described above have been designed in 2010-2011. They are well known and have been described earlier by GBIF [7] and others. Ho wever, there are many unique features in the Digitarum process. First, digitisation and collection management have been separated. Thi s should simplify the production, and has led to the choice of using just XML-documents for data management, as the solution and process are independent of the collection management software used by customers. Indeed, a relational database might not be an ideal solution at all for natural history collections, as there are few transactions. On the other hand, there is a need to keep track of the history of specimen curation, identifications, georeferencing, publishing events, which can be easily handled with a version control system such as SVN (Apache Subversion; http://subversion.apache.org/), but is not a typical feature of a relational database. XML- based document management makes it possible to easily go back to original material and retain all older versions, if needed.

Second, all material is imaged. With th e low cost of storage this is becoming increasingly popular also elsewhere. This reduces the need for handling the specimens.

Thirdly, and more importantly, comprehensive imaging makes it possible to distribute data entry and subsequent steps of the process to distance workers. Th is way costs can be reduced and access to remote experts gained for purposes such as handwriting recognition, languages, and species identification.

Fourth, comprehensive digitisation may reduce the need to access the specimens physically. Thi s is still somewhat controversial, but many studies can be carried out just using the digital copy of the specimen. When digitised, the specimens and entire collections can be stored in a re mote repository in a less expensive town and building than the big museums in city centres typically can provide. Digitarium offers this repository service for the material it digitises.

Implementation of the process described here is by no means completed. The process is being tested and refined, but is not yet very effective for large scale production. Scaling up of capacity will happen gradually, but it is too early to estimate what level of efficiency will be achieved.

ACKNOWLEDGMENTS

We thank Mikko Heikkinen for support for the URI tagging and Tapani Lahti for the ideas concerning XML-based data management. We are grateful of the help and cooperation of Greg Riccardi, Deb Paul, and others of th e Morphbank team at Florida State University. We also thank an unknown referee for improvements of th e text. This work has been financed by the European Social Fund and European Regional Development Fund.

REFERENCES

[1] Berendsohn, W.G., Chavan, V. & Macklin, J. 2010. Summary of Recommendations of the GBIF Task Group on the Global Strategy and Action Plan for the Digitisation of Natural History Collections. Biodiversity Informatics, Vol 7, No. 2.

[2] Pelkonen, V-P., Saarenmaa, H. & Laurenne, N. (editors) 2009. Luonnontieteellisten museokokoelmien digitointi. Strategia ja toimintasuunnitelma 2010-2015. Helsingin yliopisto, Luonnontieteellinen keskusmuseo 31.12.2009.

[3] Saarenmaa, H. 2009. Luonnontieteellisen keskusmuseon tietoaineistojen avoimuuspolitiikan toteuttaminen Hatikassa ja muissa aineistoissa. Luonnos, versio 4.3 http://gbif.fi/fi/node/32

[4] Haapala, J. & Lehtonen, J. Darwin Core 2009-09-23: Käyttöohje biologisten aineistojen tallentamiseen. Manuscript 2011-05-24, 26 pages.

[5] Morphbank :: Biological Imaging (http://www.morphbank.net/, 31 May 2011). Florida State University, Department of Scientific Computing, Tallahassee, FL 32306-4026 USA.

[6] Berkley, C.; Jones, M.; Bojilova, J.; Higgins, D.; 2001. Metacat: a schema-independent XML d atabase system. Scientific and Statistical Database Management, 2001. SSDBM 2001. Proceedings. Thirteenth International Conference on , vol., no., pp.171-179, 2001.

[7] Global Biodiversity Information Facility. 2008. GBIF Training Manual 1: Digitisation of History Collections Data, version 1.0. Copenhagen: Global Biodiversity Information Facility.

Using Semantic Metadata for Discovery and Integration of Heterogeneous Ecological Data

Ben Leinfelder[1], Shawn Bowers[2], Margaret O'Brien[3], Matthew B. Jones[1], Mark Schildhauer[1]

[1] NCEAS, University of California Santa Barbara
[2] Dept. of Computer Science, Gonzaga University
[3] Marine Science Institute, University of California Santa Barbara
{leinfelder, jones, schild}@nceas.ucsb.edu, bowers@gonzaga.edu, mob@msi.ucsb.edu

***Abstract*—Effective discovery and integration of ecological data within data management systems requires rich semantic information that can describe and relate the types of information contained within disparate data sets. Within the Semtools project, we have developed approaches for expressing and representing semantic annotations of data sets for supplementing attribute and data-level metadata with terms drawn from domain-specific ontologies. Annotations provide a formal mechanism that can be used together with reasoning systems to enhance existing data discovery and integration approaches. We describe extensions to the Ecological Metadata Language (EML) and associated tools for storing and using semantic annotations. Specifically, we describe new user interface components implemented within the Morpho metadata editor for capturing user-supplied semantic annotations, extensions to the Metacat system for storing and accessing annotations and corresponding OWL-DL ontologies, and a new API within Metacat that uses annotation metadata to provide concept-based search and integration of data sets.**

***Keywords*—ontologies; annotation; data discovery and integration**

I. Introduction

A major challenge in environmental information management concerns providing effective approaches for the discovery and integration of heterogeneous data sets. For instance, locating and combining relevant observational data are often critical and time-consuming steps for researchers studying phenomena at broad spatial, temporal, and biological scales [1], [2]. The underlying data sets used within such studies frequently differ in subtle and complex ways, due in part to the protocols used for data collection, the types of observations made, and the experimental and other contextual information associated with the data set. These differences in turn can lead to structural and semantic heterogeneity among data sets that make them hard to discover using current data management approaches and require considerable manual effort by researchers needing to combine data sets.

A number of recent efforts within the earth and environmental informatics communities are adopting the notion of an *observation* as a key modeling concept for enabling improved discovery and integration of scientific data [3]–[7]. These approaches provide higher-level observational data models for describing and representing observations and measurements found in underlying data sets by defining common "core" concepts such as the entities or features being observed, measurement units and protocols, and context relationships between observations [3], [7]. A major goal of these approaches is to enable interoperability and uniform access to data by abstracting away the underlying representation details that often impede integration across scientific data sets.

In this paper we describe extensions to the Ecological Metadata Language (EML) [8] and supporting tools for enabling improved discovery and integration of ecological data sets. Our work is based on the Extensible Observations Ontology (OBOE) [7], [9], which represents a generic observational model implemented in OWL-DL [10] for describing domain-specific observation and measurement types. Our approach adds additional metadata in the form of semantic annotations that link attributes within data sets to OBOE terms for describing the implicit observation and measurement types found within data sets. Semantic annotations are executable in the sense that they can be used to convert a data set into a collection of observation and measurement instances, providing a more uniform representation for expressing queries and performing integration. To support the creation of annotations, we have extended the Morpho metadata editor [11] with a high-level user interface as well as the Metacat data catalog [12] for storing and querying annotations through a new Semantic Mediation API. This API can also be used to perform basic data-level integration tasks using our prior work on the EML Data Manager Library [13].

The rest of this paper is organized as follows. Sec. II briefly describes the various components used within our approach including the extensions we have developed for Morpho and Metacat to support semantic annotation. Sec. III describes the types of data discovery queries and integration services supported by our framework. Sec. IV briefly describes related work, and we summarize our contributions in Sec. V.

II. Semtools Framework

The Semtools project has focused development efforts on three main components: a Java library to access and manipulate OBOE ontologies and semantic annotations, an annotation plugin for the Morpho metadata editor, and query extensions for the Metacat data catalog. Below we briefly describe the

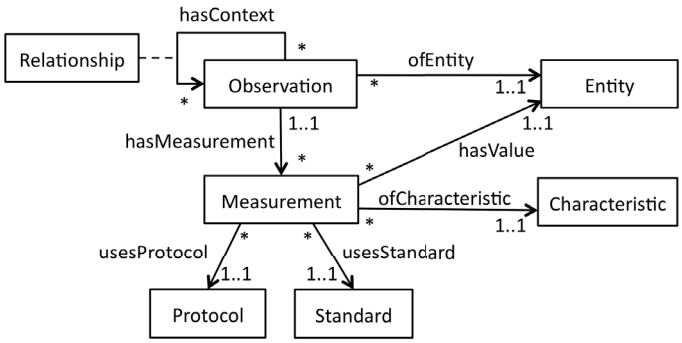


Fig. 1. Main classes and properties of the extensible observation ontology (OBOE). While shown using the Unified Modeling Language (UML), the model is defined as an OWL-DL ontology.

OBOE model, the semantic annotation approach used by Semtools, and the corresponding software components. For a more in-depth presentation of OBOE see [7], [9].

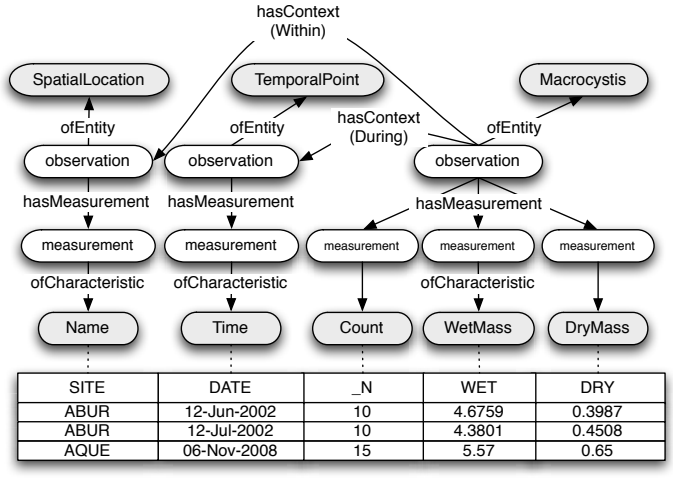
A. The OBOE Observational Model

Fig. 1 shows the main modeling constructs of OBOE (see: http://ecoinformatics.org/oboe/oboe.1.0/oboe-core.owl). An *observation* is made of an *entity* (e.g., biological organisms, geographic locations, environmental features) and serves to group a set of measurements together to form a single "*observation event*". A *measurement* assigns a value to a *characteristic* of the observed entity (e.g., the weight of a plant), and can also include *standards* (e.g., units as well as standards for coded values) and collection *protocols*. An observation can occur within the surrounding *context* of other observations (e.g., as part of a temporal or spatial context), and context may include a named relationship (e.g., "partOf", "within") that existed during the observation event. A key feature of OBOE is that it allows properties (characteristics and relationships) of entities to be asserted without being interpreted as *inherently* (i.e., always) true of the entity. Depending on the context in which the entity was observed or how the measurements were performed, an entity's properties may take on different values. OBOE allows RDF-style assertions about entities to be contextualized, and thus different values can be assigned for the same entity under distinct contexts, which is a crucial feature for modeling ecological as well as many other types of scientific data [6], [7]. In addition, OBOE is currently implemented as an OWL-DL ontology that can be easily used with (or extended by) other ontologies for specifying domain-specific types of entities, characteristics, measurement standards, protocols, and relationships. For instance, the Semtools project has defined specific OBOE extensions in collaboration with the Santa Barbara Coastal Long-Term Ecological Research Project as well as through ongoing collaborations with other projects, and general extensions exist for OBOE that define a number of common entities, measurement units, and corresponding physical characteristics.

B. Semantic Annotations

A semantic annotation consists of two parts: (i) a "configuration" of the observation model containing the specific entities, characteristics, observations, measurements, and so on (drawn from one or more domain ontologies) that appropriately capture the semantics of the data set; and (ii) a mapping between the attributes in the data set to specific measurements defined in the model configuration. Fig. 2 shows a high-level example of an annotation defined for a simple Kelp sampling data set. Here, the data set consists of five attributes (bottom of Fig. 2). Each attribute is mapped to a specific measurement type (where only the characteristic of each measurement type is shown), and measurement types are organized into observations of specific Kelp entities (shown of type "Macrocystis"), temporal points (denoted by date-times), and spatial locations (given as site names). Each measurement associated with a Kelp observation is assumed to have occurred within the site and during the given time as specified by the context relationships.

Semantic annotations can be used to facilitate discovery and integration of heterogeneous data sets. For instance, combining semantic annotations with OBOE, it is possible to discover data sets based on searches expressed over types of observations and measurements of interest. As simple examples, users can pose queries such as "*find all data sets containing observations of Kelp*" and "*find all data sets containing Mass measurements of Kelp*". Both of these queries would return the example data set in Fig. 2 since the attribute `WET` is linked to a WetMass measurement for observations of Macrocystis (where WetMass is defined as a special kind, or *subclass* of Mass, and Macrocystis is defined as a subclass of Kelp). Using semantic annotations in this way can help to increase both query recall and precision over standard keyword-based approaches [14]. In particular, by defining terms as subclasses of other terms (e.g., Macrocystis as a subclass of Kelp), term expansion can be used to increase the number (recall) of data sets returned (where subclasses of query terms are also searched). The precision of the result can be improved since queries may specify



SITE	DATE	_N	WET	DRY
ABUR	12-Jun-2002	10	4.6759	0.3987
ABUR	12-Jul-2002	10	4.3801	0.4508
AQUE	06-Nov-2008	15	5.57	0.65

Fig. 2. Partial OBOE semantic annotation for Kelp sampling data. Shaded nodes represent ontological concepts; rectangular nodes are data table attributes mapped to OBOE measurement characteristics.

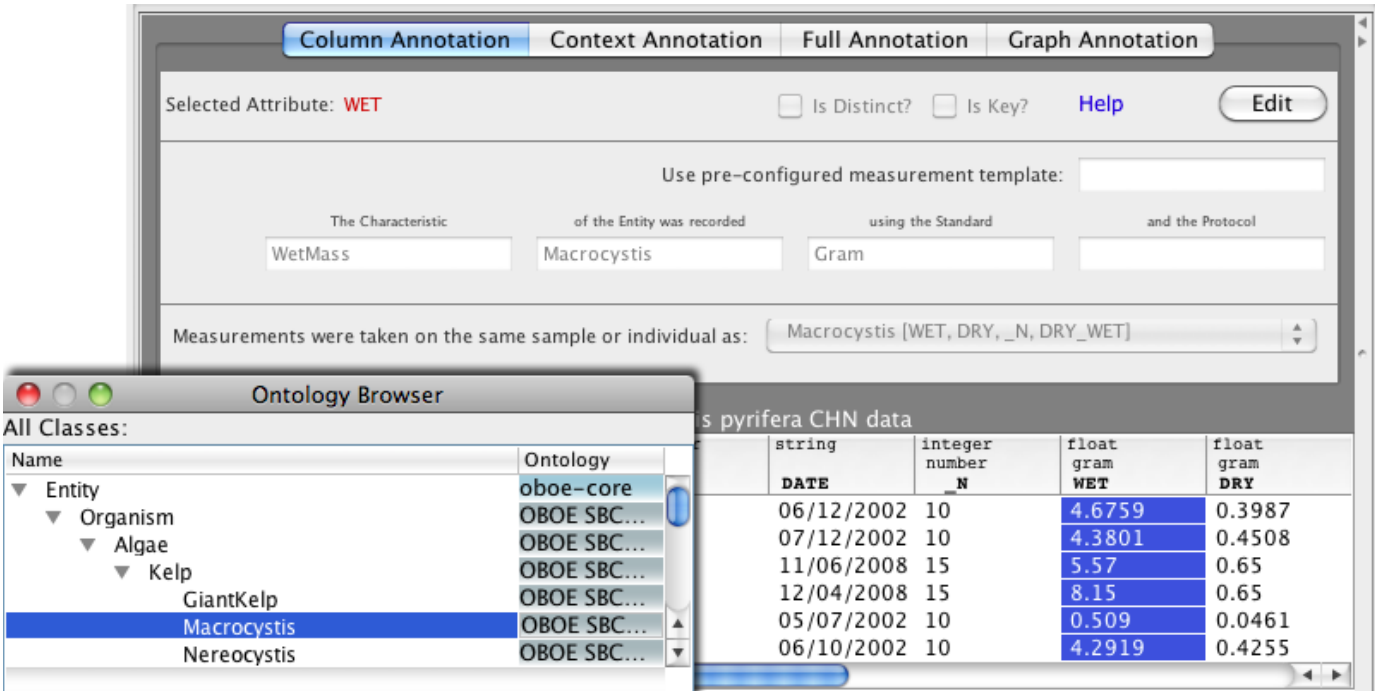


Fig. 3. Morpho metadata editor with Semantic plugin. The fill-in-the-blank interface uses natural language descriptions for intuitive editing. A searchable, hierarchical browser is used to select concepts from domain-specific ontologies.

the desired connections between terms (e.g., measurements of Mass for Kelp observations) as opposed to returning all data sets that simply mention the terms but without any explicit connections (i.e., where Mass was measured, but not for Kelp samples). Annotations also help facilitate integration by allowing tools to align data set attributes based on their declared measurement and observation types.

In general, semantic annotations provide a formal description of attribute semantics, whereas in many commonly-used metadata formats for describing data sets, only informal text-based descriptions of data attributes are permitted. In the case of EML, there is some overlap between these two mechanisms—particularly with respect to measurement standards—however, semantic annotations extend this approach by providing a general mechanism to formally associate concepts drawn from domain ontologies to attributes. In the Semtools framework, we employ an XML serialization syntax for semantic annotations that is compatible with EML but that is stored separately from the EML documentation of a data set (allowing, e.g., annotations to be used independently of EML or with other metadata standards if needed). In addition, semantic annotations can be used to "materialize" a given data set into a set of triples conforming to the model configuration given in the annotation. In other words, a tabular data set such as the one shown in Fig. 2 can automatically be converted into a corresponding collection of observation and measurement instances. This in turn enables a simple form of structural integration, where instead of having a large number of different tabular data structures, all data is represented using the standard set of structures defined by the OBOE model (see Fig. 1). Thus, materializing a data set in this way provides a more uniform structural representation that can make a number of discovery and integration tasks easier. For instance, materialization can be used to increase query expressivity by allowing searches of the form "*find all data sets containing Mass measurements of Kelp with values less than or equal to 5 grams*", which in our example can be answered by generating (i.e., materializing) the measurements associated with the `WET` and `DRY` attributes in the data set of Fig. 2.

C. The Semantic Mediation API

The Semantic Mediation API includes basic ontology management features, annotation manipulation capabilities, and simple concept navigation and visualization components. The API is intended to be a centralized toolkit for use in multiple application contexts (on either client or server deployments). The Semantic Mediation API uses both the OWL API [15] for ontology management services including ontology parsing, serialization, and simple class and property exploration as well as the Pellet description-logic reasoner [16] for classification and exposing inferred axioms in source ontologies. The inference services exposed through the Semantic Mediation API are used in both discovery and integration features described below. In our current Morpho and Metacat extensions, semantic annotations are managed and stored automatically in an underlying, local relational database. While it is also possible to use in-memory approaches for storing and querying annotations, we found the overhead to be prohibitive when large numbers of data sets are managed.

D. The Morpho Editor Plugin

The semantic-annotation editor plugin for Morpho provides a front-end to the Semantic Mediation API and allows data owners and curators to define annotations for existing EML data descriptions. The editor provides a simple "fill-in-the-blank" style form-based interface with a searchable hierarchical concept selection widget (see Fig. 3). The plugin seamlessly integrates with a standard Morpho installation and provides semantic query capabilities for locating data packages, marking up data sets within a package using semantic annotations, and saving annotations locally or to a shared repository where they can be discovered and explored by other users. The annotation editor in Morpho allows a user to view the data set being annotated as they fill in (by selecting an appropriate ontology term) the characteristic, measurement standard, protocol, and associated entity for each data set attribute. Users can also specify whether an observation spans multiple columns, and can provide context relationships between attributes (i.e., observations). The editor provides a number of additional features including the ability to view the entire annotation (similar to Fig. 2) and to specify additional mapping constraints for observations and measurements.

E. Metacat Query Extensions

The semantic plugin for Metacat augments Metacat's existing metadata storage and search by allowing annotations to be saved and queried alongside the metadata and data that they annotate. In addition to traditional keyword and spatial search criteria, the Metacat plugin allows semantic criteria to be included where they may either increase query recall using term-expansion (i.e., traversing the class subsumption hierarchy) or refine the result set by limiting matches to data sets that contain the specified observational model (e.g., combinations of OBOE-compatible entity, characteristic, measurement standard, or protocol concepts). The observational model can be leveraged further by materializing the annotation and data artifact (via the Data Manager Library [13]) into a fully instantiated OBOE model and inspecting (and querying over) the observational values themselves.

III. Discovery and Integration

In this section we describe the new data discovery and integration applications we have built using the components described above as part of the Semtools project.

A. Concept Query

The semantic query interface (see Fig. 4) is implemented as a Web application over Metacat that primarily supports locating data sets by how well their observational models match the given criteria. The interface provides *structured* as opposed to *unstructured*, i.e., keyword-based queries. In particular, query criteria given by users largely mirror the structure of a semantic annotation in that combinations of Entity, Characteristic, and Protocol are specified and optionally compounded when increased precision is sought.

As discussed above, by leveraging the relationships defined and inferred from the ontology we are able to increase recall beyond what is possible for simple keyword-based searches [14]. Broad queries return direct matches as well as subclass matches. The queries can be quickly refined when using the Web application by allowing rapid exploration of the data repository without having to define complete observational queries *de novo*. The interface allows users to specify individual classes of a measurement as well as pre-configured measurement types (representing standard data set attribute types) as defined in OBOE compatible ontologies to enable a single concept to proxy its constituent parts, namely the characteristics of particular entities that can be measured with a set of protocols and standards. This short-hand query generation can save users time in specifying their queries, and highlights a compelling reason for using OBOE extension ontologies. Measurement templates can also be leveraged when creating or editing semantic annotations in the Morpho interface.

Using compound semantic query criteria applies a finer-grained filter on the data sets that are returned. Results can be restricted to only those data sets that include measurements for a set of specific characteristics of a particular observational entity. Furthermore, a query can specify that those measurements come from precisely the *same instance of that entity*; a feature that fully exercises the comprehensive observational structure expressed in the annotation and enables higher query precision as described above.

B. Data-Level Query

For even more precise recall, the OBOE model can be (partially) materialized (see above) during the query stage after which a data range filter can be applied. Different techniques are available for merging the annotation with the data that it describes, but for performance reasons a hybrid approach has been adopted in which preliminary search results from a structured query are collated and only that subset is materialized. Because our corpus is described using EML in conjunction with the annotation syntax, the Data Manager Library [13] is used to load the described raw data (into a relational database) while the annotation informs the correct use of the Data Manager query and filtering features. For any measurements that match the concept query criteria, we verify that those measurements (e.g., attributes) contain data values within the range specified in the initial semantic data query and return the data packages that contain them (see Fig. 4).

C. Data Integration

The materialization routine for semantic data queries can help in enabling data integration. In addition to inspecting data for values within a range and returning the data sets that contain a match, the data integration feature of the Semtools Web application goes further by constructing a synthetic data product (table) that represents the complete results of the query in terms of both the attributes and the values that are returned. Each original data set may have very different syntactic structures (e.g., column number, naming, order) but could share a

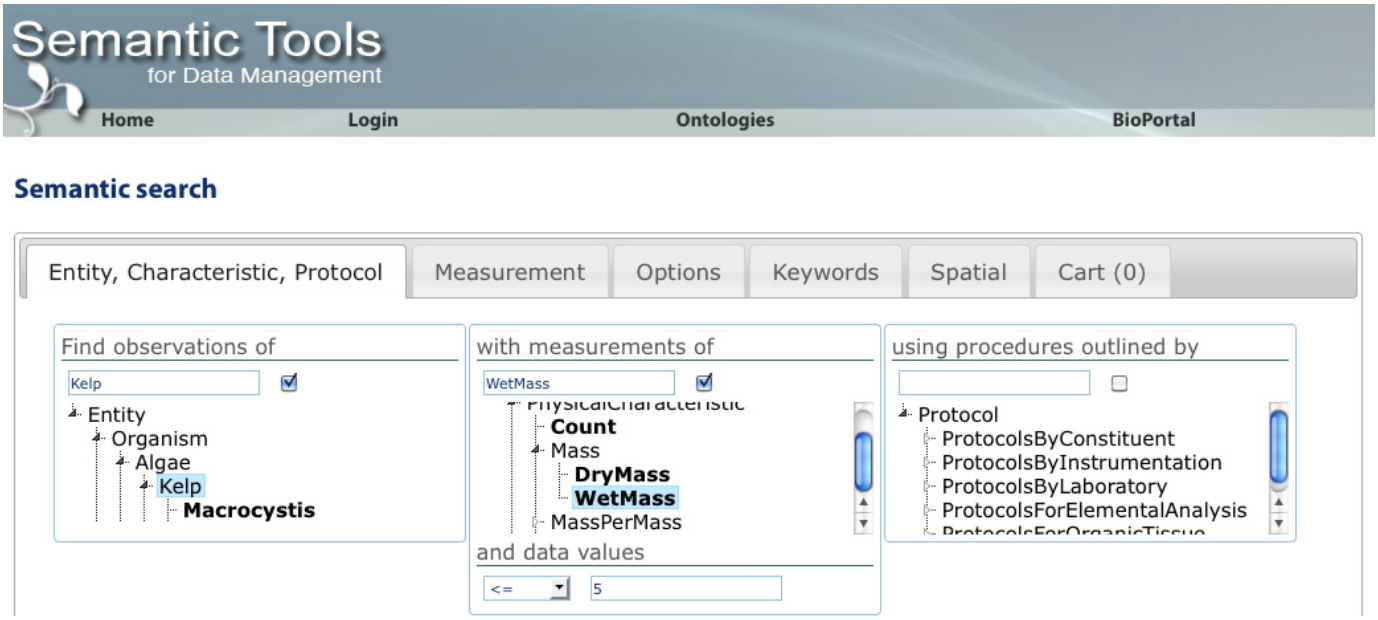


Fig. 4. Semantic data query web interface. Data packages containing observations of Kelp Wet Mass less than or equal to 5 [grams] are returned. Additional search options and compound query criteria can be specified within the other tabs. Matches can be saved in the data cart for later exploration.

subset of attributes that are semantically compatible as defined in accompanying annotations. These compatible attributes can then become the basis for a synthetic data set. Fig. 5 illustrates the data integration support provided in the current implementation of the Web application. Consider the two data packages (denoted A and B) in Fig. 5. Annotations (denoted C and D) are used to determine semantically equivalent data attributes contained in the data sets (denoted by E and F). The attributes `plot` and `site` are considered equivalent measurements of the characteristic Location; attributes `weight` and `wt` both map to the same characteristic Mass. The Semantic Mediation API computes an equivalence among attributes based on their corresponding annotations. The Data Manager Library is then used to load the data sets and then query each data set to produce and merge a synthetic result data set.

While this approach provides a preliminary form of data-level integration, we are currently developing additional algorithms for determining compatibility of annotated measurements (e.g., to include unit information such as that gram and ounce are both mass units) and for converting measurement values using ontologically-defined unit conversions (e.g., 1000 milligrams in a gram), which will further support automated data integration through the Web application.

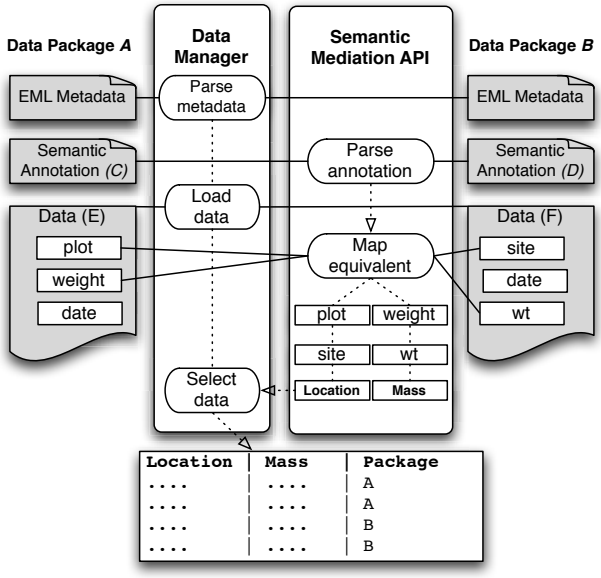


Fig. 5. Integration query across multiple data packages (A, B). Annotations (C, D) determine semantically equivalent data attributes contained in the data objects (E, F). Attributes `plot` and `site` are considered equivalent measurements of the characteristic Location; attributes `weight` and `wt` both map to the same characteristic Mass. The Semantic Mediation API utilizes the Data Manager Library to load and query the source data informed by semantic similarities between the structurally disparate data attributes.

IV. Related Work

The need for more semantic mechanisms to describe observational data has led to many proposals for observational data models (e.g., [3], [5], [17]) and ontologies (e.g., [4], [6], [18]). The work presented here is complementary to these efforts by providing a concrete set of software components that have been integrated with popular metadata tools (namely, Metacat [11] and Morpho [12]) to provide a more uniform, semantic view of heterogeneous observational data. By extending Morpho and Metacat to support semantic annotations, these tools can provide additional help to researchers interested in performing synthetic studies by providing semantically-enhanced discovery and integration services, which are largely lacking in many existing environmental information manage-

ment frameworks [19].

Our work on using semantic annotations for data integration is closely aligned to traditional information integration approaches (e.g., [20]), where a global mediated schema is used to (physically or logically) merge the structures of heterogeneous data sources using mapping constraints among the source and target schemas. As such, the observational model we employ in our framework can be viewed as a (general-purpose) mediation schema for observational data sets. This schema can be augmented with logic rules (as target constraints) where semantic annotations are used as mapping constraints. However, instead of users specifying logic constraints directly, we provide a high-level annotation language and user-interface components (through Morpho) that can simplify the specification of mappings and more naturally aligns with the observation model.

Annotations are playing a more prominent role in database systems, e.g., the MONDRIAN system [21] employs an annotation model and a set of query operators to manipulate both data and annotations. However, users must be familiar with the underlying data structures (schemas) to take advantage of these operators, which is generally not feasible for observational data in which data sets exhibit a high degree of structural and semantic heterogeneity. Our annotation approach used to extend EML is also similar in spirit to a number of other high-level mapping languages used for data exchange (e.g., [22], [23]). Our approach differs by being specifically tailored to the OBOE observational model, which in turn simplifies the annotation language, making it in general easier for users to specify annotations for observational data. Our approach also provides well-defined and unambiguous mappings from data sets to the observation and measurement model, which is critical for providing automated, high-quality data integration services over heterogeneous observational data.

V. Conclusion

The Semtools project has been successful in exploring and codifying technologies and techniques for applying semantic concepts to observational data. By providing mechanisms for linking data sets to ontological terms organized in a high-level observational model (e.g., OBOE), these new extensions to Metacat and Morpho help to overcome a number of limitations in existing metadata management systems that strive to provide effective data discovery and integration features. Our close involvement with the SONet Project (Scientific Observations Network) [24] encourages continued use-case refinement that will inform future semantic tool development and place an emphasis on intuitive interfaces and incremental adoption. This varied community of stakeholders is firmly invested in the use of cutting edge semantic solutions that will ultimately benefit multiple science disciplines by reducing obstacles to broad data sharing and innovative reuse.

Acknowledgement

This work supported in part through NSF grants 0743429 and 0753144.

References

[1] B. Worm, E. Barbier, N. Beaumont, J. Duffy, C. Folke, B. Halpern, J. Jackson, H. Lotze, F. Micheli, S. Palumbi, E. Sala, K. Selkoe, J. Stachowicz, and R. Watson, "Impacts of biodiversity loss on ocean ecosystem services," *Science*, vol. 314, no. 5800, pp. 787–790, 2006.

[2] S. Pennings, C. Clark, E. Cleland, S. Collins, L. Gough, K. Gross, D. Milchunas, and K. Suding, "Do individual plant species show predictable responses to nitrogen addition across multiple experiments?" *Oikos*, vol. 110, no. 3, pp. 547–555, 2005.

[3] OGC, "The OpenGIS Observations and Measurements Encoding Standard (O&M)," http://www.opengeospatial.org/standards/om.

[4] P. Fox, D. McGuinness, L. Cinquini, P. West, J. Garcia, J. Benedict, and D. Middleton, "Ontology-supported scientific data frameworks: The virtual solar-terrestrial observatory experience," *Computers & Geosciences*, vol. 35, no. 4, pp. 724–738, 2009.

[5] D. Tarboton, J. Horsburgh, and D. Maidment, "CUAHSI community observations data model (ODM), version 1.0 design specifications," http://water.usu.edu/cuahsi/odm/.

[6] C. Mungall, "Representing phenotypes in OWL," in *Proc. of the Workshop on OWL: Experiences and Directions (OWLED)*, 2007.

[7] S. Bowers, J. Madin, and M. Schildhauer, "A conceptual modeling framework for expressing observational data semantics," in *Proc. of the Intl. Conf. on Conceptual Modeling (ER)*, 2008, pp. 41–54.

[8] E. Fegraus, S. Andelman, M. Jones, and M. Schildhauer, "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, 2005.

[9] J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa, "An ontology for describing and synthesizing ecological observation data," *Ecological Informatics*, vol. 2, no. 3, pp. 279–296, 2007.

[10] "OWL DL," http://www.w3.org/TR/owl2-overview/.

[11] D. Higgins, C. Berkley, and M. Jones, "Managing heterogeneous ecological data using Morpho," in *Proc. of the Intl. Conf. on Scientific and Statistical Database Management (SSDBM)*, 2002, pp. 69–76.

[12] C. Berkley, M. Jones, J. Bojilova, and D. Higgins, "Metacat: A schema-independent XML database system," in *Proc. of the Intl. Conf. on Scientific and Statistical Database Management (SSDBM)*, 2001, pp. 171–179.

[13] B. Leinfelder, J. Tao, D. Costa, M. Jones, M. Servilla, M. O'Brien, and C. Burt, "A metadata-driven approach to loading and querying heterogeneous scientific data," *Ecological Informatics*, vol. 5, no. 1, pp. 3–8, 2010.

[14] C. Berkley, S. Bowers, M. Jones, J. Madin, and M. Schildhauer, "Improving data discovery for metadata repositories through semantic search," in *Proc. of the Intl. Conf. on Complex, Intelligent and Software Intensive Systems (CISIS)*, 2009, pp. 1152–1159.

[15] "The OWL API," http://owlapi.sourceforge.net/.

[16] Clark and Parisa, "Pellet: OWL 2 Reasoner for Java," http://clarkparsia.com/pellet/.

[17] Unidata, "network Common Data Form (netCDF)," http://www.unidata.ucar.edu/software/netcdf/.

[18] "Semantic Web for Earth and Environmental Terminology (SWEET)," http://sweet.jpl.nasa.gov/sweet/.

[19] M. Jones, M. Schildhauer, O. Reichman, and S. Bowers, "The new bioinformatics: Integrating ecological data from the gene to the biosphere," *Annual Review of Ecology Evolution and Systematics*, vol. 37, pp. 519–544, 2006.

[20] P. Kolaitis, "Schema mappings, data exchange, and metadata management," in *Symposium on Principles of Database Systems (PODS)*, 2005, pp. 61–75.

[21] F. Geerts, A. Kementsietsidis, and D. Milano, "Mondrian: Annotating and querying databases through colors and blocks," in *Proc. of the Intl. Conf. on Data Engineering (ICDE)*, 2006.

[22] R. Fagin, L. Haas, M. Hernandez, R. Miller, L. Popa, and Y. Velegrakis, "Clio: Schema mapping creation and data exchange," in *Conceptual Modeling: Foundations and Applications*, 2009, pp. 198–236.

[23] Y. An, J. Mylopoulos, and A. Borgida, "Building semantic mappings from databases to ontologies," in *Proc. of the AAAI*, 2006.

[24] "SONet: Scientific Observations Network," http://sonet.ecoinformatics.org/.

Provenance and Quality Control in Sensor Networks

Barbara Lerner[1], Emery Boose[2], Leon Osterweil[3], Aaron Ellison[2], Lori Clarke[3]
[1] Mount Holyoke College
[2] Harvard University
[3] University of Massachusetts Amherst
blerner@mtholyoke.edu, boose@fas.harvard.edu, ljo@cs.umass.edu,
aellison@fas.harvard.edu, clarke@cs.umass.edu

***Abstract*—Scientists and society increasingly rely on streaming data from electronic sensors to assess, model, and forecast environmental changes. Because analyses of time-series data require uninterrupted data streams or datasets, scientists regularly fill gaps in the data by substituting modeled values. As modeling increases in complexity, the provenance metadata needed to describe and define processes used to model data and create derived datasets quickly exceeds the capacity of individual flags or groups of flags to annotate individual data values. In theory, necessary provenance metadata could be captured in narrative form, but the time and effort required to do so are prohibitive. A system that can capture provenance metadata automatically and allow scientists to query them for useful details is what scientists really need. In this paper we describe a system that uses Little-JIL, a process programming language, to rigorously define modeling and data-derivation processes, and a mathematical graph structure – a Data Derivation Graph (DDG) – that precisely describes execution histories. Our system and approach support understanding the (potentially) different processes used to create data values, reasoning about the soundness of these processes, and helping to ensure that the data processing in sensor networks is reliable and reproducible.**

Keywords—provenance metadata, scientific workflow, sensor network, Little-JIL

I. INTRODUCTION

Scientists and society increasingly rely on streaming data from electronic sensors to assess current environmental states and to forecast future environmental changes. Because analyses of time-series data require uninterrupted data streams or datasets (i.e., there must be a reliable observation for each time slot), scientists regularly fill gaps or correct "problems" in data streams by substituting modeled values for missing, out-of-range, or suspect observations. Different scientists substitute, model, or gap-fill data differently, and some approaches can be inconsistent with subsequent analyses. Such inconsistencies can undermine the quality and reduce the reliability of derived datasets, but these changes in quality and reliability often are invisible to subsequent users of the derived datasets. Therefore, it is critically important to be able to identify which data values represent actual observations and which have been modeled, and how modeled values have been computed. Furthermore, even observed values may undergo subsequent revision; e.g., to compensate for sensor drift that is discovered at a later time. Finally, a given data value may have been adjusted more than once. All of this suggests that the different data items in a dataset should be annotated with information (*metadata*) about exactly how their values were derived. A full history of all of the adjustments to a given datum is referred to as the data item's *provenance*; the annotation is referred to as *provenance metadata*.

Often scientists "flag" values in a dataset using schemes that identify special conditions attendant to the data. At the Harvard Forest Long Term Ecological Research (LTER) site, current practice is to flag estimated values (including modeled values) with the single letter "E." But a simple flag (or even several flags) is insufficient to answer all of the questions that may arise with regard to data provenance. For example, if a precipitation datum in a dataset actually originated at another site (e.g. due to sensor failure), it may be important to know which site was the origin of the datum, especially if it turns out that the second site was also experiencing sensor reliability problems on that date. Or if measurements are corrected *post-hoc* (e.g. to compensate for sensor drift), we may need to know how the data were corrected and over what range of dates, in order to correctly update derivative data products (e.g. monthly or annual summaries). Finally, if a datum was computed (not actually observed) using a model, it is important to track software and modeling tools used, as there can be variation in precision and accuracy, for example, among the different versions of the tools and algorithms used in model computation.

As data modeling increases in complexity, the provenance metadata needed to describe and define the processes, models, and associated derived data rapidly exceeds the expressive power of modest numbers of individual flags or groups of flags. Provenance metadata can be captured in narrative form, but the considerable effort required to capture these metadata accurately and then to decipher them correctly renders narratives and their analysis error-prone, especially since narratives are rarely machine readable. A system that can capture provenance metadata automatically and allow scientists to query them for useful details is what scientists really need. Our solution is to continually record comprehensive metadata as the data are collected and processed so that scientists can (re)examine the data, perhaps in ways that were not anticipated, or not possible, initially. In this paper we describe our experience in treating scientific data values to be the outputs of

the execution of a (scientific data processing) process where the provenance metadata of the generated data is a summary of the execution history of the process. Our work uses Little-JIL, a process programming language, to define such processes, and a graph structure, called a Data Derivation Graph (DDG), to summarize their execution histories. The rigorous definitions and semantics of Little-JIL, and of the derived DDGs,

In this paper we propose an extension of the current approach that will combine (1) automated processing of real-time measurements, along with gap filling for missing or out-of-range values, and (2) user-initiated post-processing to correct for sensor drift and update modeled values using both preceding and subsequent measurements.

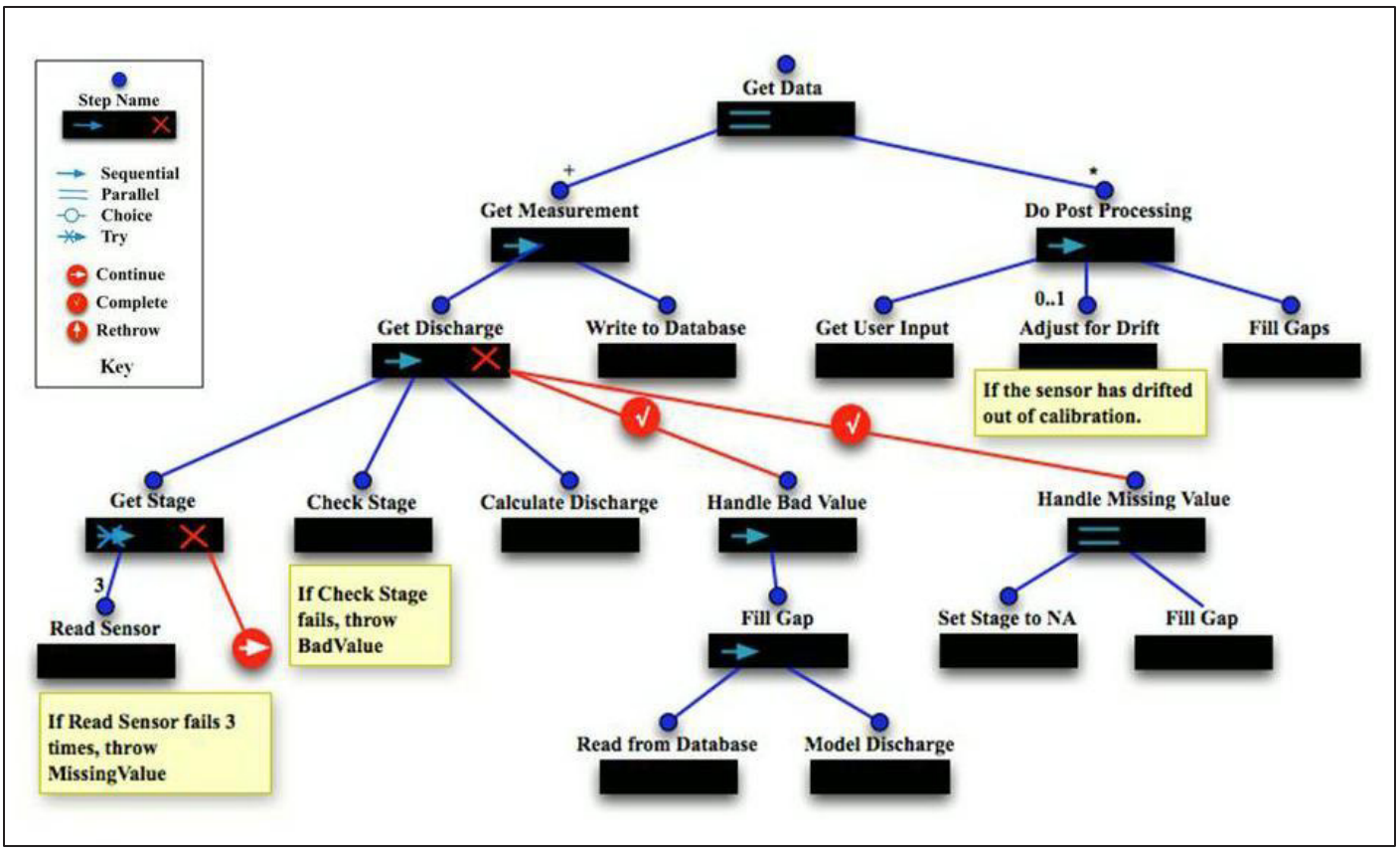


Figure 1. Little-JIL diagram for the stream discharge process.

support reasoning about the processes used to build data and datasets. This can build confidence in, and ensure the quality of, scientific data and derived data products [1].

II. STREAM GAGE EXAMPLE

Our example is an ongoing study of water movement through small forested watersheds at the Harvard Forest. The study relies on automated measurements of stream discharge (rate of flow) at a series of stream gages. At each gage, a pressure sensor is used to measure the stage or height of the water at the gage. A datalogger samples the sensor every 10 seconds, then calculates and retains 15-minute averages. The 15-minute values are retrieved from the datalogger, checked to see if they are within range, and (if they are) used to calculate stream discharge based on empirical flow equations for the particular gage. The resulting time-stamped 15-minute values of discharge are then posted online (http://harvardforest.fas.harvard.edu:8080/exist/xquery/data.xq?id=hf070).

III. PROVENANCE AND LITTLE-JIL

Little-JIL [2,3,4] is a graphical process programming language that supports the representation and execution of processes that may involve the interaction of multiple agents to accomplish a task (note: our terminology differs somewhat from that used in the Open Provenance Model [5]; in particular, the OPM concept of "process" corresponds more closely to the Little-JIL concept of "step"). Little-JIL processes are defined using a hierarchical decomposition of steps and substeps. This hierarchical decomposition allows a process to be viewed at various levels of abstraction, with a step's substep structure defining the way in which the step is to be carried out. A leaf step is carried out by assigning it to an "agent", an entity that is responsible for assuring the acceptable performance of the step, but in a way that is outside of the direct control of Little-JIL. Agents may be either humans or automated devices (e.g. software systems or sensors).

Artifacts flow through a Little-JIL process by being passed as parameters between steps and substeps. Each edge in a Little-JIL diagram carries a specification of the artifacts that are being passed between parent and child, along with binding information needed to relate the data flowing along an edge to the parameter specifications of the steps that are connected by the edge. Little-JIL edges can also carry cardinality information that specifies the number of instances of the substep that are to be instantiated for execution. The cardinality specification may be an integer or a Boolean expression used to determine the circumstances under which the substep is to be generated for execution. To simplify the depiction, the Little-JIL diagram does not directly show the artifacts, but a user can see this information by clicking on an edge in the Little-JIL editor.

Each step also specifies the resources required for the step to execute (the step's agent is considered to be a resource, but additional resources may also be specified), any exceptions that may be thrown by the step, and any provisions that the step may make for handling exceptions that could be thrown by any of the step's descendants.

The graphical representation of a Little-JIL step with its different badges and possible connections to other steps is shown in the key to Figure 1. The interface badge is a circle on the top of the step name that connects a step to its parent. The interface badge contains the specification of any artifacts that are either required for, or generated by, the step's execution as well as the type of the agent required to execute the step. Below the circle is the step's name. The icon at the left of the black rectangle identifies the sequencing construct that controls how the step's substeps are executed. There are four possibilities: sequential (all substeps in order from left to right), parallel (all substeps in any order or concurrently), choice (choose one substep at runtime), and try (execute substeps from left to right until one succeeds). The red X at the right edge of the black rectangle attaches a step to its exception handlers. Exceptions may be "thrown" by any of the descendants of a step. Control flow then goes to the nearest ancestor with a handler for that exception. After completing execution, the handler determines where execution should resume. There are three possibilities: continue (continue the step following the substep that threw the exception), complete (treat the parent step of the handler as having completed its execution and continue from there), and rethrow (throw the same exception thereby passing the exception up the step hierarchy to the next ancestor with a handler for that exception).

Figure 1 shows the Little-JIL diagram for the stream discharge process. The parallel root step (Get Data) builds and updates a database of sensor data through the concurrent operations of its two substeps, Get Measurement and Do Post Processing. Get Measurement collects and processes sensor data in real time and adds a record to the database for each measurement. Under normal conditions Read Sensor returns a measured value, Check Stage checks to see that the value is in range, Calculate Discharge calculates stream discharge, and the resulting values are added to the database. Exceptional conditions are handled by the corresponding exception handler. For example, if Check Stage determines that the measured value is out of range, the Handle Bad Value step generates a modeled discharge value based on preceding measurements read from the database. Similarly, if Read Sensor fails on three attempts, the Handle Missing Value step assigns a value of NA to stage and concurrently generates a modeled value for discharge.

Meanwhile Do Post Processing (shown here in abbreviated form) runs concurrently with Get Measurement. In contrast to Get Measurement, which runs continuously to process streaming data in real-time, Do Post Processing only executes infrequently, when a scientist determines that post processing is required. Do Post Processing first gets input from the user (including the range of dates and adjustment and modeling parameters), optionally adjusts a block of measurements for sensor drift, and then updates all modeled values in that block using both preceding and subsequent data.

We attach cardinality to substep edges to control the number of times that a step is repeated. In this example, the edge to Get Measurement has a cardinality labeled "+", meaning that the step is done one or more times. The edge to the Do Post Processing step has a cardinality labeled "*", meaning that the step is done zero or more times. The edge to Adjust for Drift has a cardinality label "0..1", meaning the step is done either 0 or 1 times, thereby making this activity optional. Finally, the edge to the Read Sensor step is labeled with a cardinality of 3, meaning that we will try to read the sensor 3 times before deciding that the sensor is unreachable. Due to the semantics of the Try step, Get Stage is complete as soon as Read Sensor successfully gets a value. If Read Sensor fails 3 times consecutively, it will throw an exception that will be handled by the Handle Missing Value exception handler attached to the Get Discharge step.

One of the strengths of Little-JIL is the ability to represent processes at any desired level of detail or abstraction. In our example, each of the leaf steps could be decomposed into its constituent substeps to show (for example) the equations used to calculate discharge from stage (Calculate Discharge) or the more complex series of calculations used to model discharge based on recent precipitation and discharge (Model Discharge). At the same time, the entire process shown here might be embedded in a much larger process that calculates water flux in a watershed by integrating measurements such as precipitation, evapotranspiration, stream discharge, water content of snow pack, soil moisture, and height of the water table.

The Little-JIL diagram provides a rigorous specification of the process but does not tell us what actually happened in any particular execution of the process. For that, we need the information contained in the DDG that is produced when a Little-JIL process is executed. Figure 2 provides examples, in the form of four DDG fragments, of different ways in which the process shown in Figure 1 can be executed, leading to the

creation of a single stream discharge value. A DDG consists of two kinds of nodes and two kinds of edges. In Figure 2, rounded nodes represent process steps that have been executed, while rectangular nodes represent values that have been used and generated by these steps. Different colors are used to denote different kinds of steps and different kinds of assignment of a missing value for stage and a modeled value for stream discharge. The last three scenarios take advantage of Little-JIL's ability to precisely describe and handle exceptions. In each case the DDG shows the exact derivation of the final stream discharge value. In particular, the bottom yellow oval in each figure represents execution of the step that writes

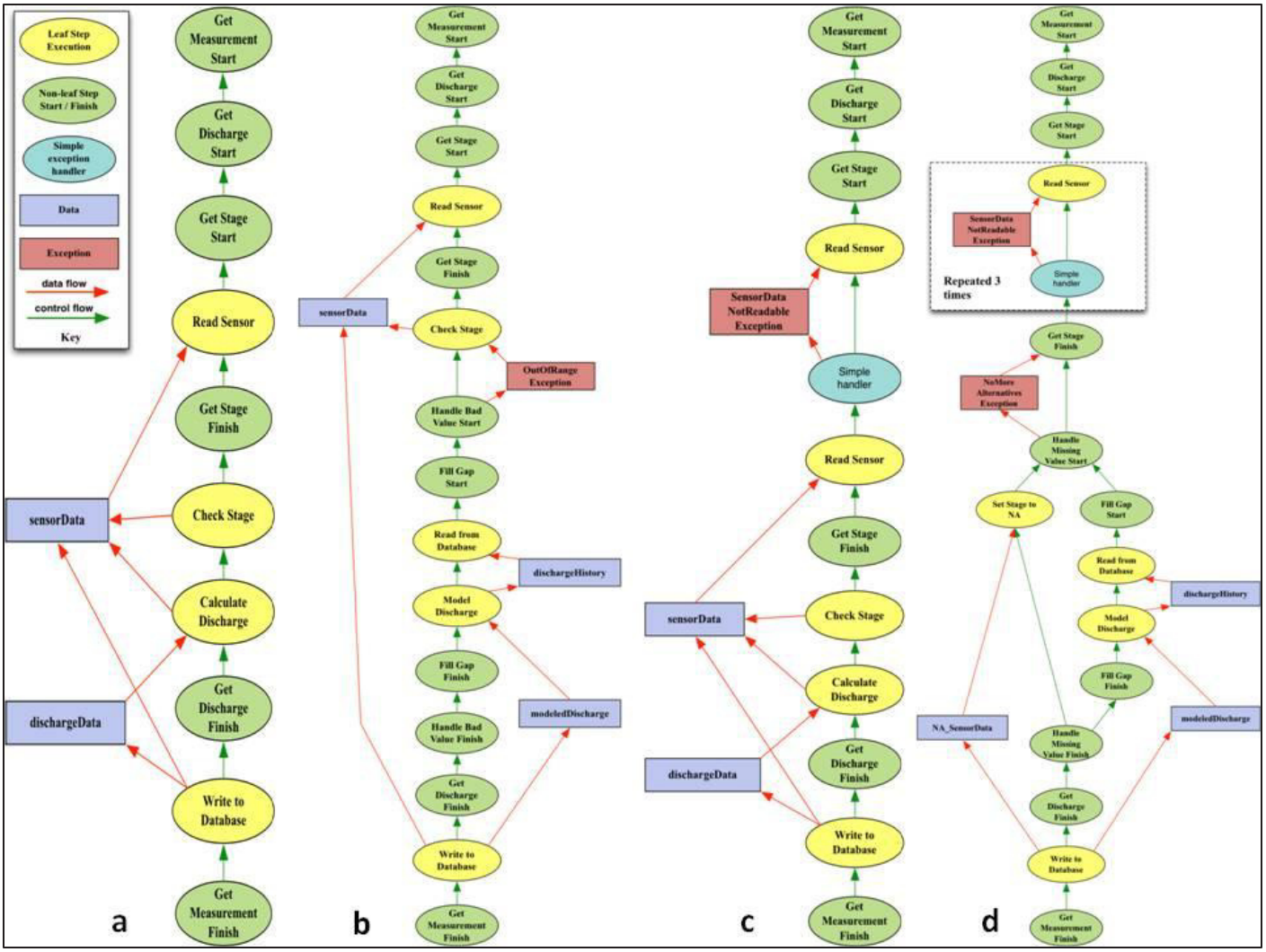


Figure 2. Four possible DDGs resulting from a single execution of the Get Measurement step: (a) normal sensor reading, (b) out-of-range value, (c) retry of Read Sensor, (d) missing value after three successive failures of Read Sensor.

values. Green edges represent the flow of control between steps while red edges show the flow of data that is generated by one step and then used as input by others.

The graphical representations in Figure 2 show the flow of data and control under four scenarios: (a) an in-range value is returned by the sensor and used to calculate stream discharge, (b) the Check Stage step determines that the sensor value is out of range and so a modeled value of stream discharge is generated, (c) the first attempt to read the sensor fails so the Read Sensor step is tried again, successfully returning a value on the second try, (d) Read Sensor is tried three times and fails to return a value on any of the three tries, resulting in the sensor data and discharge data to the archival database. By following the red arrows up from this oval, the scientist can observe the origin or provenance of each value that is saved in the database. In the first and third cases, the observed sensor value and corresponding calculated discharge value are saved. In the second case, the observed sensor value is saved and a modeled discharge value is generated and saved since the observed sensor value is not usable. In the fourth case, a special NA (missing) value is recorded for the sensor value along with the modeled discharge value.

Most of the processing demonstrated in this example is sequential, leading to a single, straight control flow path

through the process. The fourth case, however, demonstrates parallel control flow that occurs during the execution of the Handle Missing Value step. Here the recording of NA for the stage value happens concurrently with the calculation of the Fill Gap step. Note that the Fill Gap step under Handle Missing Value is a reference to the same collection of steps that is rooted at Fill Gap under the Handle Bad Value exception handler. This ability to refer to steps defined elsewhere in the process provides the ability to duplicate the same behavior in different contexts throughout a process, where the context is determined by the parameter values passed in for use by the step.

IV. Related Work

Scientific data provenance is receiving increased attention [6, 7]. The Open Provenance Model [5] defines a graph representation of provenance metadata, similar in many respects to the DDGs presented here. One area of future work is to map DDGs into OPM to allow interoperability with other provenance repositories.

One significant difference between Little-JIL and other scientific workflow approaches is in exception handling. Exception handling constructs were introduced into programming languages, such as C++ and Java, to help deal with erroneous or unlikely situations where the appropriate response is often best determined in the calling scope of where the exceptional situation arose. In Little-JIL, the hierarchical levels of the process definition serve as scopes that are searched upward for an exception handler. This provides the benefits that normally come from exception handling mechanisms, most importantly, the ability to cleanly separate exception handling code from code describing the computations to be carried out in nominal (usually expected) cases, avoiding the spaghetti code that otherwise frequently arises when code to handle exceptional cases is interleaved with the processing of nominal cases.

Some workflow management systems provide support for detecting failures during execution, such as the failure of a web service, and offer a limited number of ways to manage those failures [8,9]. Kepler [10] provides the ability to annotate a collection with an exception, which an actor can then use to filter out collections that contain exceptions. User-defined exception handling is just beginning to appear in scientific workflow languages [11,12,13].

In addition to the ability to define complex exception handling, the provenance recorded in DDGs distinguishes exception objects from other types of data. We expect that a common concern among scientists is to be able to easily identify when the execution of a process encountered problems. By explicitly capturing this information in a DDG, it will be easier for scientists to perform queries that will identify the problems encountered during process execution. In the sample DDGs shown in this paper, we distinguish exception nodes by their color. As we develop the query mechanisms to access information from DDGs, we plan to give the scientist the ability to develop queries that can distinguish exceptional situations from expected situations as well.

Provenance metadata has previously been used to track changes made as sensor data is republished [14]. The emphasis in that work has been on linking together sites on the Internet that are using each other's data in order to track how the data are republished and to control access to the data. Thus, provenance metadata are used to track how sensor data are accessed and updated even though they may be distributed widely. The focus of our work is on using provenance information to support reasoning aimed at assuring that processes have the desired properties of correctness, robustness, and access control, and also to allow processes to be used directly in computing the data itself, as in the post-processing work described earlier.

V. Discussion and Future Work

Our experience to date suggests that our approach is effective in capturing detailed and accurate provenance information. Moreover, our approach supports the capture of execution details down to low levels, if those low level details are incorporated into the Little-JIL process definition. However, DDGs quickly can become large and unwieldy, as can be seen even in our simple example. We are now investigating ways to store DDGs using various database technologies that support querying and visualization. Such databases will allow scientists to focus on particular areas of interest, such as data collected from a specific instrument at a specific site on a specific date. Because many data items follow the same path through the process, we are exploring database representations that allow us to compress the stored representation considerably, yet allow us to extract provenance metadata of an individual datum without paying the storage cost of the complete DDG. Even in our simple example (e.g. Figure 2d), a repeating node pattern is easily identified. Other kinds of repetition can arise in a DDG that represents identical derivation paths of different individual discharge values. However, in more complex processes, individual paths may diverge, especially if different data values use different computations, if there is parallelism in computation, or if data values often require special error handling. A similar compression approach has been pursued by Anand et al. [15].

We are also investigating visualization mechanisms [16, 17, 18] that build upon queries of the provenance metadata to streamline the amount of data presented to the scientist. As mentioned earlier, one of the strengths of Little-JIL is the way in which the hierarchical decomposition of processes allows processes to be viewed at varying levels of abstraction. The DDGs that we produce capture the complete data flow, via the red edges, but also maintain information about the hierarchy expressed in the process, via the non-leaf start and finish nodes. We plan to take advantage of this information in visualization, to allow the scientist to zoom in and out on provenance detail, and also allow the scientist to express queries at varying levels of abstraction, again as reflected in the process. For example, the substeps rooted at Get Discharge could be

collapsed into a single node showing only the stage and discharge values output by the step or fully expanded to show intervening details (Get Discharge Start to Get Discharge Finish, as shown in Figure 2).

ACKNOWLEDGMENTS

This work was supported by NSF grants DBI-1003938, CCF-0905530, CCR-0205575 and IIS-0705772, and is a contribution from the Harvard Forest Long-Term Ecological Research (LTER) program. We would like to thank Margo Seltzer for her contributions to the preliminary database work. We would also like to thank the students and programmers who have contributed to the creation of various versions of the stream discharge process and who have worked on the development of the software to capture DDGs: Alexander Wise, Cori Teshera-Sterne, Morgan Vigil and Sofiya Taskova.

REFERENCES

[1] E. R. Boose, A. M. Ellison, L. J. Osterweil, R. Podorozhny, L. Clarke, A. Wise, J. L. Hadley, and D. R. Foster. 2007. Ensuring reliable datasets for environmental models and forecasts. Ecological Informatics 2: 237-247.

[2] A. Wise. Little-JIL 1.5 language report. Technical report, Department of Computer Science, University of Massachusetts, Amherst, MA 01003, October 2006.

[3] A. Wise, A. G. Cass, B. S. Lerner, E. K. McCall and L. J. Osterweil, Using Little-JIL to coordinate agents in software engineering. In Proceedings of the Automated Software Engineering Conference, pages 155-163, Grenoble, France, September 2000.

[4] B. S. Lerner. Verifying process models built using parameterized state machines. In 2004 ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA '04), pages 274–284, Boston, MA, July 2004.

[5] L. Moreau, B. Plale, S. Miles, C. Goble, P. Missier, R. Barga, Y. Simmhan, J. Futrelle, R. E. McGrath, J. Myers, P. Paulson, S. Bowers, B. Ludäscher, N. Kwasnikowska, J. V. den Bussche, T. Elkvist, J. Freire, and P. Groth. The open provenance model (v1.01, http://eprints.ecs.soton.ac.uk/16148/1/opm-v1.01.pdf).

[6] S. B. Davidson and J. Freire. Provenance and scientific workflows: challenges and opportunities. In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, pages 1345–1350, Vancouver, June 2008. ACM.

[7] Y. L. Simmhan, B. Plale, and D. Gannon. A survey of data provenance in e-science. SIGMOD Record, 34:31–36, September 2005.

[8] A. Azimi and S. Parsa. A reliable framework for adaptive scientific workflow management systems based on SOA. In Proceedings of the 13th International Conference on Advanced Communication Technology (ICACT 2011), pages 1358–1363, Seoul, 2011.

[9] Q. L, W. Lin, W. Dou, J. Jiang, and J. Chen. A QoS-aware exception handling method in scientific workflow execution. Concurrency and Computation: Practice and Experience, 23, in press.

[10] T. M. McPhillips and S. Bowers. An approach for pipelining nested collections in scientific workflows. SIGMOD Record, 34:12–17, September 2005.

[11] J. Li, Y. Mai, and G. Butler. Implementing exception handling policies for workflow management system. In Proceedings of the Tenth Asia-Pacific Software Engineering Conference (APSEC '03), Los Alamitos, CA, USA, 2003. IEEE Computer Society.

[12] R. Tolosana-Calasanz, J. A. Bañares, O. F. Rana, P. Álvarez, J. Ezpeleta, and A. Hoheisel. Adaptive exception handling for scientific workflows. Concurrency and Computation: Practice and Experience,,, 22:617–642, April 2010.

[13] X. Fei and S. Lu. A dataflow-based scientific workflow composition framework. IEEE Transactions on Services Computing, 99(PrePrints), 2010.[13

[14] U. Park and J. Heidemann. Provenance in sensornet republishing. In IPAW, pages 280–292, 2008.

[15] M. K. Anand, S. Bowers, T. McPhillips, and B. Ludäscher. Efficient provenance storage over nested data collections. In Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, EDBT '09, pages 958–969, New York, NY, USA, 2009. ACM.

[16] M. K. Anand, S. Bowers, and B. Ludäscher. A navigation model for exploring scientific workflow provenance graphs. In Proceedings of the 4th Workshop on Workflows in Support of Large-Scale Science, WORKS '09, pages 2:1–2:10, New York, NY, USA, 2009. ACM.

[17] P. Macko and M. Seltzer. Provenance Map Orbiter: Interactive exploration of large provenance graphs. In Proceedings of TaPP '11, 3rd Usenix Workshop on the Theory and Practice of Provenance, Crete, Greece, June 2011.

[18] O. Biton, S. Cohen-Boulakia, S. B. Davidson, and C. S. Hara. Querying and managing provenance through user views in scientific workflows. In ICDE '08: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, pages 1072–1081, Cancun, Mexico, April 2008. IEEE Computer Society.

Interactive Visualization of Spatial and Temporal Patterns of Diversity and Abundance in Ecological Data

Tuan Pham[1], Steven Highland[1], Ronald Metoyer[1], Donald Henshaw[2], Jeff Miller[1], Julia Jones[1]
[1] Andrews LTER, Oregon State University, Corvallis, Oregon
[2] US Forest Service, Pacific Northwest Research Station, Corvallis, Oregon
pham@eecs.oregonstate.edu, highlans@geo.oregonstate.edu, metoyer@eecs.oregonstate.edu, dhenshaw@fs.fed.us, jeffrey.miller@oregonstate.edu, jonesj@geo.oregonstate.edu

***Abstract*—Analysis of spatial and temporal patterns of diversity and abundance in ecological data has been an important focus in ecology. Nevertheless, ecological data such as multi-species data sets are often difficult to analyze because species are usually unevenly represented and multiple environmental covariates may describe their distributions. Although typical univariate, bivariate, and multivariate statistics provide rigorous tests of hypotheses, they have limited capacity to quickly identify relationships among multiple species and environmental covariates, or detect change over time. We propose a novel visualization technique, the Diversity Map, which facilitates the visual inspection of the distribution, abundance, and covariates of large multi-species data sets using an interactive web-based visual interface. To develop this tool, we have taken a user-centered design approach, in which our team of ecologists, information managers, and computer scientists collaborate closely during the development process. Initial findings indicate that this tool is extremely valuable for ecologists in the early stages of data exploration, prior to further statistical analysis. In this paper, we discuss our design approach, the design elements, and implementation of the Diversity Map tool and we demonstrate how the tool can help scientists gain insights into spatial and temporal patterns of ecological data. The use of this tool is illustrated with data on moth diversity and abundance from the HJ Andrews Experimental Forest.**

Keywords—interactive data visualization; web-based application; multivariate data; user-centered design; moth diversity and abundance; HJ Andrews Forest

I. INTRODUCTION

Understanding how spatial and temporal patterns of species diversity and abundance respond to environmental gradients and temperature are fundamental problems in ecology. For example, ecologists hypothesize that the emergence, abundance, and distributions of moths may be indicators of phenology and its effects in mountain landscapes as well as of broader biological diversity in plant types and physical environments [1, 2]. Therefore, the conservation of moths, especially rare moths, may depend on the conservation of associated vegetation habitat [3].

A common approach to verifying these hypotheses is to collect data and then utilize statistical tests to draw conclusions. In addition, recent developments in statistics and data mining have resulted in methods to describe patterns and make predictions automatically [4]. These approaches work well when the number of testing variables is small and/or hypotheses are preconceived. Otherwise, a more comprehensive approach may be to enable ecologists to directly explore the data, form hypotheses, and discuss their findings with others, prior to specific hypothesis testing. Interactive visualizations of the data offer the potential to allow this kind of exploration, if the representation can reveal patterns and/or trends across variables. While typical static charts such as scatter plots and histograms have traditionally been utilized by ecologists to explore diversity and abundance patterns, little work has been done to develop interactive visualizations that support multivariate multi-species data.

Before we introduce our visualization tool, consider our particular ecological problem of studying diversity and abundance of moths. Ecologists have sampled moths in the 64-km^2 H.J. Andrews Experimental Forest (HJA) and Long Term Ecological Research (LTER) site within the Willamette National Forest, Lane County, Oregon. Moths were sampled at 20 sites every two weeks from May-October from 2004 to 2008. The data set has been difficult to analyze because the data set is large (>69,000 individual moths), many species (>500) are present, common species are widespread, and most species are rare (see Section II.A). Typical univariate and bivariate statistics utilized by ecologists have limited capacity to identify relationships among species and environmental covariates, or detect change over time in such complex multivariate datasets. For example, a tremendous amount of information is concealed in diversity indices (e.g. Shannon Index [5, 6]); regressions limit researchers to species-by-species tests; and some multivariate methods have limited tests of species-environment relationships. Yet while exploration of single variables (attributes) via static histograms is useful (or rank/abundance curves [6], in particular as shown in Fig. 1), these approaches are visually overwhelming when a large number of variables and/or subsets of data are involved.

Visualizations may assist in the process of data exploration and manipulation, and serve as a complement to statistical approaches. From the computing perspective, the moth data set

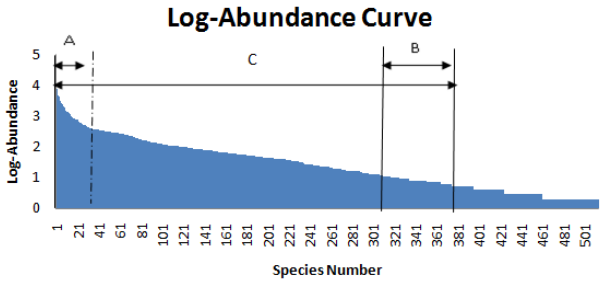


Figure 1. Log Abundance curve showing the distribution of moth species in the moth dataset. 'A' shows the common moths, 'B' shows the rare moths, and 'C' shows the common through rare moths.

presents 1) a challenging large multivariate data set visualization problem, 2) a unique visual exploration process that involves inspecting distributions and relationships of distributions as opposed to specific data samples, and 3) valuable supporting materials for sharing of scientific findings, if the representation of the data is readily available.

In our research, we have developed a novel visualization technique, the Diversity Map (DM) [9], that facilitates the visual inspection of the diversity, abundance, and relationships among multiple variables using an interactive web-based visual interface. To develop the tool, we have taken the user-centered design approach in which ecologists work closely with computer scientists during all stages of the design process [10], [11]. Initial findings from the application of the tool to the HJA moth data set indicate that it is highly valuable for ecologists in the early stages of data exploration and collaboration. In particular, ecologists can use this tool to quickly form an overview of their entire data, drill down to subsets of data, detect relationships among variables, identify and share hypotheses for further exploration, and download subsets of data for standard statistical analysis. Moreover, since the tool is web-based and readily available, it may potentially target a broader user pool, including educators and students.

II. METHODS

We have developed the DM tool based on the *information visualization reference model* [7, 8], a widely-used software architecture pattern that models the visualization process as discrete steps from collecting the source data and transforming them to appropriate formats to mapping data to visual representations and ultimately supporting view transformation via user interactions (Fig. 2). The outcome of the process is an interactive visualization that helps users complete their tasks and/or gain additional insights into their data. In addition to utilizing this model, we have integrated the users (ecologists) into the design process with the user-centered design approach

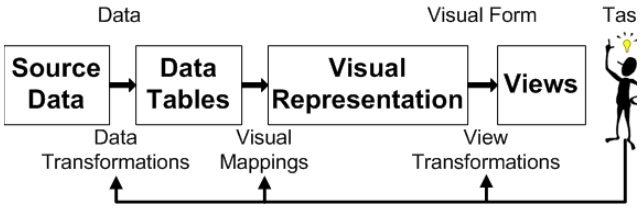


Figure 2. Information Visualization Reference Model [7, 8] illustrating the steps involved in building an interactive visualization.

[10, 11]. This section describes data sets and the steps involved in development of the tool.

A. Source Data – Moth Trapping

Moths were collected at 20 locations in the Andrews Forest (Fig. 3) 10 times per year during the summers of 2004 to 2008 (2-week sampling periods), using UV light traps. Moth abundance refers to the number of individuals caught in a single trap in a single night, or the total number of individuals in any aggregated assemblage of trapping events. Host plants for moths, if known, were based on Miller and Hammond [12]. Additionally, the following environmental variables were used to explain the distributional patterns of moths: calendar day (sampling period), temperature (accumulated heat-units), vegetation type, watershed, and elevation. Values of vegetation type, watershed, and elevation are determined based on trap sites and values of temperature are based on sampling periods.

In summary, a total of 69,168 individual moths from 514 species were captured (Fig. 1). Species richness was high, but most species were rare, producing highly varied patterns of diversity (Fig. 1). Fifty-four (10%) of the 514 moth species were represented by only 1 individual, and 46 (9%) were represented by 2 individuals.

We used two subsets of the entire moth dataset in the analyses: 26 common moth species and 66 rare moth species. We define common moth species (n=26) as those for which 500 or more individuals were captured over the entire five-year sampling period. We define rare moth species (n=66) as those for which a total of 5-10 individuals were captured over the five year sampling period. Note that we do not include moths with 1-4 individuals as part of the rare moths because we assume that an average abundance of at least one per year will provide enough information to identify the moth's spatial and temporal associations. Moth species with 1-4 individuals will not provide the level of detail needed to sufficiently identify the environmental associations of the moth species. For example, singletons and doubletons are very difficult to understand because they do not occur often enough to analyze statistically.

The 26 most common moth species ('A' in Fig. 1) accounted for 41,889 individuals (60.6% of the total abundance). The 66 moth species considered as rare ('B' in Fig. 1) accounted for 467 individuals (0.7% of the total abundance).

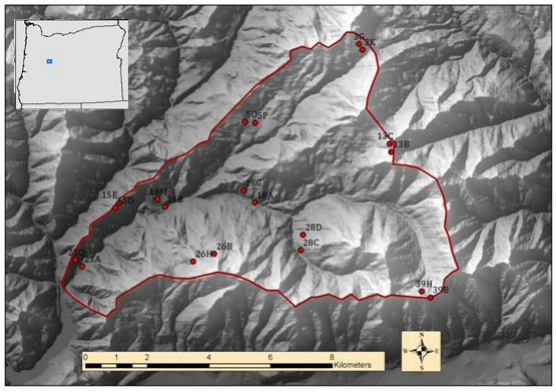


Figure 3. Map showing the location of the Andrews Forest in the central western Cascades, Oregon with 20 moth trap sites (red dots). The red line is the boundary of the forest.

B. *Data Transformation*

We compiled the common and rare moth data sets into a table format, with each column corresponding to an attribute (variable) and each row corresponding to a sampled moth species. Specifically, each row represents a moth species with non-zero individual abundance collected at a trap site on a sampling date. We augment each sampled species with the aforementioned environmental variables. The structure of the data set is described in Table I. Note that the DM representation, which we describe in the next section, is currently designed to visualize only categorical data. We transform quantitative attributes into categorical attributes by discretizing or binning values into ranges.

C. *Visual Mappings – The Diversity Map Representation*

The DM representation is based loosely on the parallel coordinates [13] and small multiple histograms techniques for visualizing multivariate data. In this representation (Fig. 4 and 5), each attribute is represented as one of a set of parallel (vertical) axes, similar to the layout of a parallel coordinates visualization. Unlike traditional parallel coordinates, however, each data object (or each sampled moth individual in the case of the moth data sets) is represented with a semi-transparent rectangle placed on each attribute axis at the discretized range corresponding to the individual's value for that particular attribute. The representation is designed primarily for categorical data, so continuous numerical attributes are discretized into bins called "buckets." The sizes and numbers of buckets for discretized continuous attributes were based on convenient divisions of the data (e.g., 100-m intervals for elevation, two-week intervals for calendar date, and 100-degree intervals for accumulated heat units).

TABLE I. STRUCTURE OF THE MOTH DATA SET

Attribute Name	Type	Description
LEP_NAME	categorical	Lepidoptera (moth) scientific name; includes genus and species
LEP_FAMILY	categorical	Lepidoptera taxonomic family
LEP_GENUS	categorical	Lepidoptera taxonomic genus
FOOD_PLANT	categorical	Host functional feeding group
TRAP_ID	categorical	Identifier for a trap site
ELEVATION	numerical	Elevation. Discretized by 100m band.
HABITAT	categorical	Habitat
WATERSHED	categorical	Watershed
COLLECT_PERIOD	categorical	2-week collect period. E.g., '7.2' represents the second half of July
COLLECT_YEAR	categorical	Collect year
TEMPERATURE	numerical	Temperature (Heat unit). Discretized by 100 unit band.
NO_INDIV	numerical	Number of individuals

We treat all individual moths equally; each semi-transparent rectangle representing one moth individual contributes an equal, fractional amount of opacity to the bucket in which it is placed. Because the range of opacity levels is limited, we scale the number of individuals in each bucket according to the total abundance of all individuals in the visualization. Thus, the opacity of each bucket x is calculated as $f(x) = |x|/|total|$, where $|x|$ denotes the number of individuals in bucket x and $|total|$ is the total number of individuals from the visualized data set. Although we use linear scaling in our implementation, the method can accommodate other forms of scaling, such as logarithmic, for species whose abundances span multiple orders of magnitude [14]. We choose white as the background color and blue as the foreground color, because the human eye is known to be more sensitive to changes in blue than in other colors [15]. We map opacity values to values in

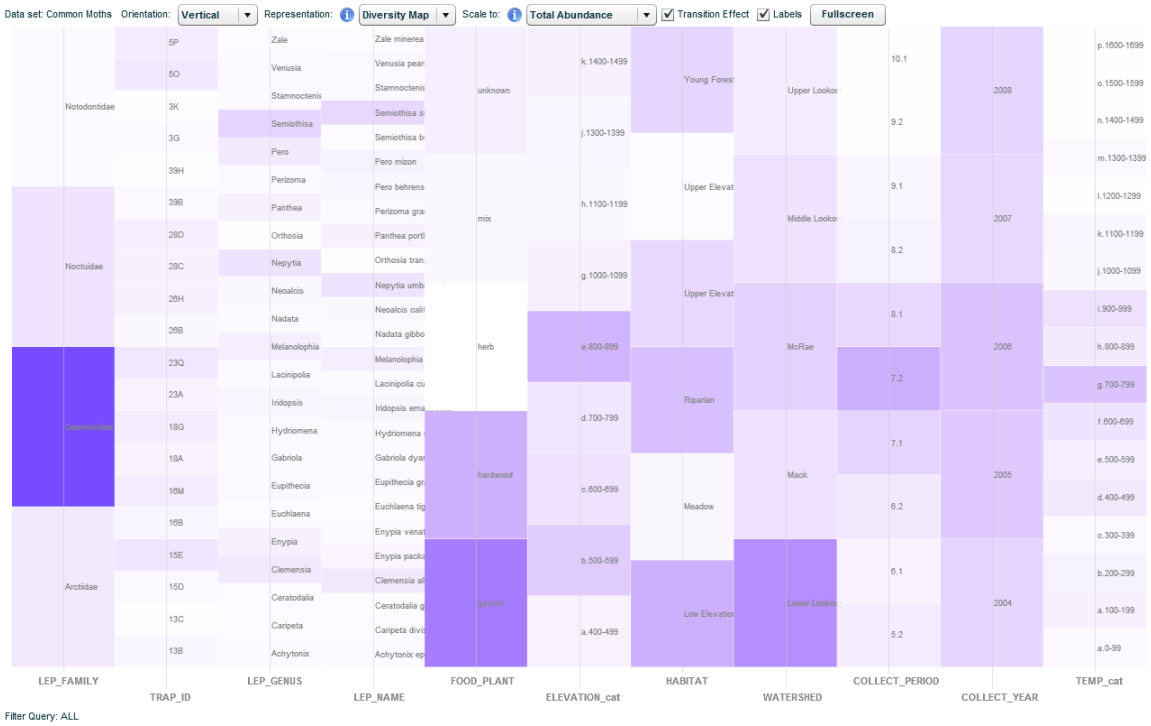


Figure 4. The DM representation of common moths. The data set contains 41,889 individual moths and 11 attributes (columns from left to right: LEP_FAMILY, TRAP ID, LEP GENUS, LEP NAME, FOOD PLANT, ELEVATION, HABITAT, WATERSHED, COLLECT PERIOD, COLLECT YEAR,

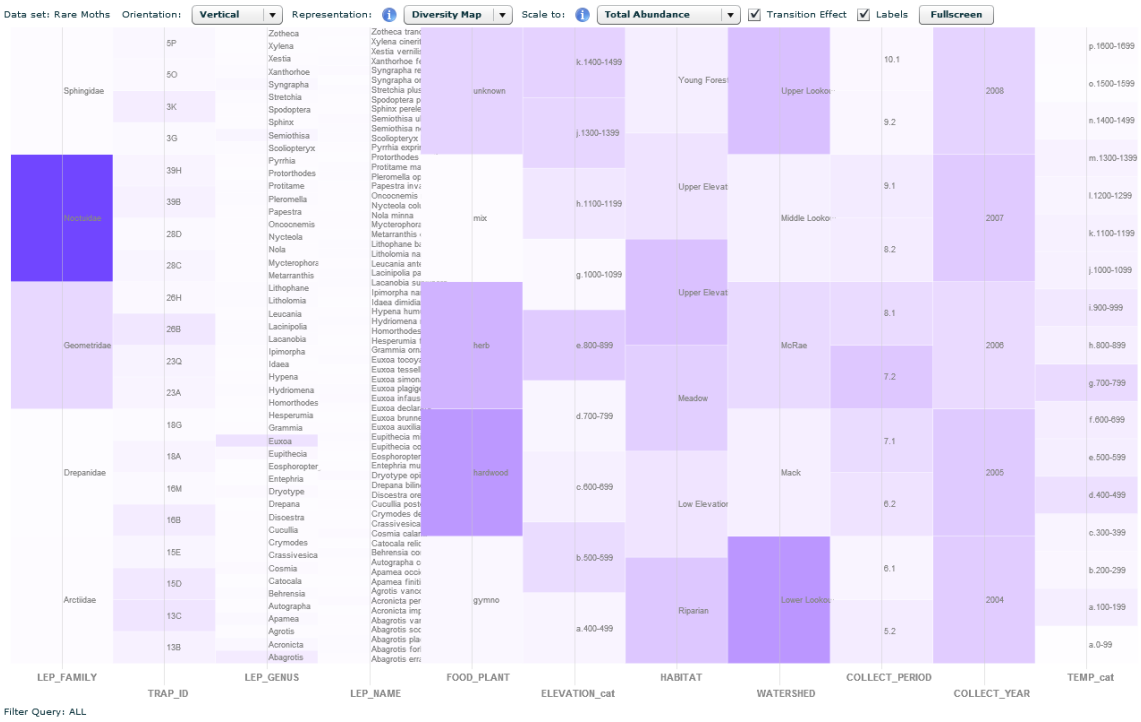


Figure 5. The DM representation of rare moths. The data set contains 467 individual moths and 11 attributes ordered as in Fig. 4

the CIELAB color space [16], which is perceptually uniform, meaning that a visual difference in color opacity is equally perceptible across the range of that color. We then convert CIELAB values to RGB values for representation on a computer screen.

Alternatively, the DM representation can be understood by imagining each attribute axis as a histogram over the values of that attribute, constructed in 3D space by stacking semi-transparent tiles on top of each other. When viewed from above, the taller stacks of tiles appear darker, while the shorter stacks appear lighter, according to the total combined contribution of the tiles in each stack to that stack's opacity. In addition to the DM representation (opacity encoding), the visualization tool also allows users to switch to a small multiple histograms representation (bar length encoding) (Fig. 6).

The DM created in this analysis expresses diversity and abundance patterns of an attribute by the number of buckets with non-zero opacity and by the color distribution across the opaque buckets of that attribute, respectively.

D. View Transformations – Interactivity

A primary characteristic that differentiates the DM tool from static charts typically employed by ecologists is that the tool supports a wide range of interactive features. These features allow the transformation of the view to alternative views so that users can interact with and explore their data. In particular, these features can be used to query the data (e.g., filtering), to change the representation of the data (e.g., switch between the Diversity Map and small multiple histograms representations, re-order the attribute axes, or sort the buckets within an attribute), or to show additional relevant information (e.g., tooltips, rich data pop-ups).

Data filtering extends the static DM to facilitate subsetting of data. For example, a user can constrain, or "filter," a single attribute or multiple attributes to one or more particular values (buckets) (e.g. show all moths that were sampled at TRAP_ID X and in COLLECT_YEAR Y) (Fig. 7). The remaining attributes then display the distribution of only those individuals that fall within the specified range of the filtered attribute values. Filtering facilitates direct comparison of the attributes of a subset of specific samples as well as comparisons of subsets of data.

Filtering is accomplished through direct manipulation of buckets. Users can simply click on a bucket to add/remove the corresponding attribute value to/from the filter. A filter 'status' bar at the bottom will show the current filter query. To construct a complex filtering query consisting of multiple buckets (or attribute values), we follow a simple and commonly used rule articulated by ecologists: buckets within an attribute are connected by the "OR" condition, whereas groups of filtered buckets across attributes are connected by the "AND" condition. Additionally, we plan to add an 'export' feature to the tool to allow users to export and download subsets of data for standard statistical analysis. To some extent, the tool can be used as a visual query builder to construct the query quickly and intuitively.

To further support comparison of attributes of interest, users are also given the ability to reorder the axes horizontally and to sort the buckets of a single attribute by abundance or by alphabetical order of value name if desired. Users can also hold the mouse pointer over a particular bucket to display the number of individuals falling into that bucket, and they can rotate the representation to accommodate their orientation preference (portrait or landscape) or their screen dimensions.

Figure 6. The small multiple histograms representation of common moths. Users can select their preferred representation in the drop-down list located on the control bar at the top.

Furthermore, the tool allows interactive identification of additional relevant information. The DM tool supports rich data pop-ups, which may display researcher-provided information on any of the buckets. For example, double-clicking on a trap ID pops up the aerial photo of that trap site in the Andrews forest (Fig. 7). Each bucket can potentially be linked to other data sources such as a GIS map, a Wikipedia page, or even another visualization.

E. Implementation

The DM tool was developed using Flex 3 and the Degrafa graphics framework. Flex 3 (available at <http://opensource.adobe.com/wiki/display/flexsdk/Download+Flex+3>) is an open-source framework by Adobe for creating Flash rich internet applications. Degrafa (available at <http://www.degrafa.org/>) is an open-source graphics framework that facilitates the process of creating pre-composed graphics in Flex 3. In particular, Degrafa helps create lightweight geometry building blocks such as rectangular buckets and attribute axes in the DM tool. Since Flash is web-based, no installation of the tool is required and it can be accessible on any browser or device that supports Flash.

In addition to the input data table as described in Section II.B, each application requires an additional metadata table that describes the valid domain for each of the visualized attributes. This metadata table enumerates all possible values for each attribute (e.g., lists each Lepidoptera family name present in the data for attribute LEP_FAMILY) and determines the default ordering for each axis. Additionally, any enumerated value in the metadata table can be augmented with other relevant data such as a URL link to an image of the actual trap indicated by TRAP_ID, or to a GIS map for any listed WATERSHED). Currently, both tables (input data and metadata) are stored in comma-separated values (CSV) format. In future work, we plan to extend the tool to load the input data and metadata directly from a database management system (DBMS), and take advantage of the highly structured metadata employed by the HJA LTER website [17] to make this tool more generic and easily applicable to other population data, such as HJA plant and birds data sets.

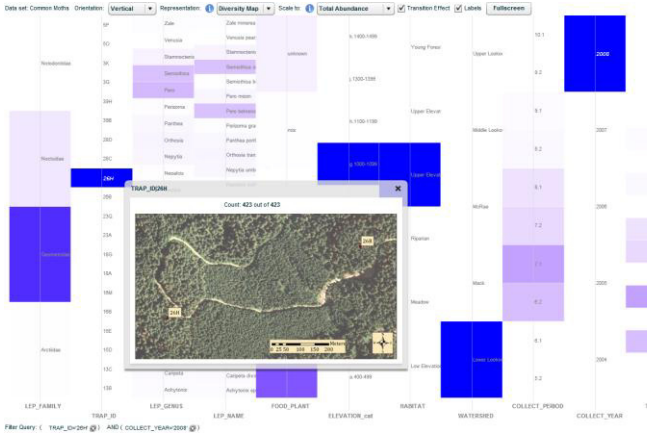


Figure 7. The DM representation of common moths sampled at TRAP_ID '26H' and in COLLECT_YEAR of '2008'. Rich data pop-up showing an aerial photo of the trap location.

F. User-Centered Design with Ecologists

A close collaborative effort between ecologists and computer scientists was required to understand the analysis process for integration of the DM into active research. We employed a user-centered, participatory design approach (Fig. 8) [10, 11] where the ecologists were included as part of the design team from the beginning of the collaborative effort. The initial prototype of the DM served as the starting point for this particular collaboration.

The initial prototype was initially developed for a small subset of the data, and it proved invaluable as a means for stimulating discussion and identifying design alternatives. In early meetings, the prototype served as a way to introduce the ecologists to the visual representation in the particular context of their data set. Subsequent meetings followed a very informative and dynamic process. In particular, each session generally started with the computer science team running the visualization, projecting the view onto a large screen for the entire team to view. The ecologists would then begin to explore the data set in an iterative fashion, asking questions and modifying views to answer those questions, and repeating. The process was typically very fast-paced and very collaborative with team members posing questions to each other and devising views together to answer those questions. When a question could not be answered using the provided

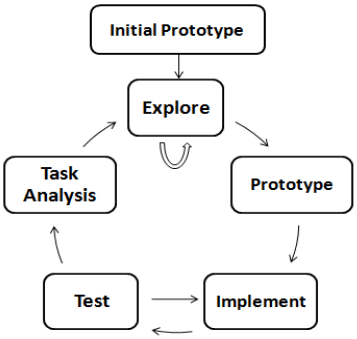


Figure 8. The collaboration between ecologists and computer scientists taking an iterative user-centered, participatory design approach

representation and interactions, the entire team would break from the exploration cycle to discuss how the system could be modified to further enhance the application. In the weeks following each meeting, the computer science team would integrate the design modifications into the system in preparation for the next design meeting. As the design matured, the work centered more on dedicated exploration and analysis of the data set.

III. Results and Discussion

In this section, we illustrate the value of the DM tool by several example scenarios of ecologists exploring the moth data sets and we discuss what we have learned from our interdisciplinary collaboration.

A. *Exploration of the moth data sets – Example scenarios*

Visualizations of common moths and rare moths can be accessed at <http://purl.oclc.org/diversitymap/commonmoth> and <http://purl.oclc.org/diversitymap/raremoth>, respectively. The ecological findings presented in this section are primarily for demonstrating the utility of the tool. Ecology readers are encouraged to refer to [18] for more detailed analysis of these findings.

First, without requiring any interactions from users, the overview of moths (Fig. 4 and 5) quickly suggests that common moths are associated with common habitats (conifer forests in the HJA) and rare moths are associated with rare habitats (meadows in the HJA). In addition, the visualization shows that common moths are mostly conifer-feeders and rare moths are mostly hardwood, herb, and grass-feeders. That is, the view of common moths (Fig. 4) shows 'gymno' is the most opaque bucket within FOOD_PLANT axis and the view of rare moths (Fig. 5) shows 'herb' and 'hardwood' are the most opaque buckets within the same axis.

Second, consider this example, which demonstrates how interactions facilitate the investigation of temporal relationships in the moth data sets. Because moth development is temperature dependent, ecologists hypothesize that adult moths emerge earlier in warm years and later in colder years. According to the temperature records, while 2004 was a warm year, 2008 was a much colder year. Ecologists can filter the moth records by COLLECT_YEAR and/or COLLECT_PERIOD to observe temporal trends. The views help verify that the peak in common moth abundance occurred earlier in 2004 (and 2006) than in 2008 (Fig. 9 left and right). Note that they show moth capture by 2-week sampling period (8th column) and by degree days (last column). In 2004, most moths were captured in sampling periods 7.2 and 8.1 with very few/no moths captured after 8.1, whereas in 2008, moths were captured in sampling periods 7.1 to 8.1 and continued to be captured until 9.1. Common moths were initially captured in a much more concentrated time span in 2004 than 2008, with many more moths initially captured later in the year in 2008 than in 2004. In this example, while ecologists need to observe only three attributes (COLLECT_YEAR, COLLECT_PERIOD, and TEMPERATURE) to answer their question, they can potentially look at other attributes for additional insights. For example, they may initially pre-define the ordering of moth species in LEP_NAME attribute (e.g., by abundance) and then quickly verify whether the ordering pattern remains consistent over these two years.

B. *User-Centered Design*

The user-centered design process was important in reaching a design that truly met the needs of the target users (ecologists). An initial prototype was a key component in starting the 'discussion' between ecologists and computer scientists and helping the design team to understand the exploration process. Although the prototype may not be the final design, some means for rapidly exploring the data allows the team members to begin to understand the typical process and types of questions they can and would like to ask of the data.

Characteristics/Process. Given interactive tools, ecologists were able to quickly and iteratively explore data that was originally in a very inaccessible format. The visualization provided an environment in which ecologists could rapidly answer questions and visually verify expected relationships. The process was typically iterative with several cycles of starting with a question, taking an exploration path, getting insight, and then starting over with a different path through the data. In some cases, ecologists felt the need to explore two paths simultaneously to observe the differences in the outcome. This multiple path exploration capability is a fundamental requirement of creativity tools [19]. Data analysis through visualization must support the creative process of hypothesis generation (Fig. 10).

Data Queries. In this particular collaborative effort, the

Figure 9. The DM representation of common moths sampled in COLLECT_YEAR of '2004' (left) and '2008' (right)

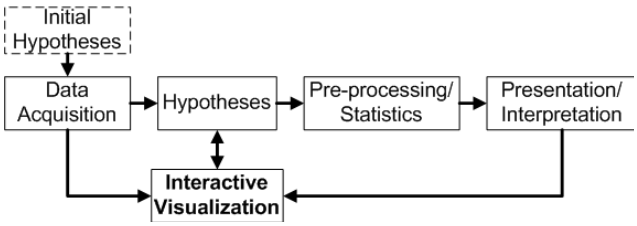


Figure 10. The visualization driven data analysis process

visualization served as a means for rapid high-level exploration of complex data that was then followed with detailed statistical analyses. Data exploration tools, such as the DM, which overview the data, should provide mechanisms for exporting subsets of data associated with the current view so that scientists can conduct appropriate statistical analyses.

Communication. On several occasions an ecologist sought to explain a particular insight or finding by walking the team through the necessary interactions to produce a specific view. Exploration tools must provide mechanisms for storing and retrieving history in order to help users tell their stories. In addition, the tools need to permit users to mark and recreate paths of exploration in order to explain ideas to one another.

Context of Collaboration. Our meetings were typically held in a conference room in the computer science building. On several occasions, the team would have benefited from being located in the context of the ecologist so that the team could refer to or use artifacts that are typically at their disposal – such as topographic maps. A more contextual design process that included, for example, sessions in the office of an ecologist or visits to field sites, might have revealed additional useful views/tools that would provide powerful insight capabilities when combined with the visual representation.

Educational Outreach. Education and outreach are key components of the H.J. Andrews Experimental Forest and LTER. We believe that visualization tools are promising in this setting, because they provide a mechanism for clearly communicating complex ideas and data through images, which are often more easily explained than data sets and scientific findings. We are currently integrating the tool into the HJA LTER website (<http://andrewsforest.oregonstate.edu/data/tools/software.cfm?topnav=149>) to make it accessible to a broader audience, including scientists, students (K-12 and undergraduate), and educators. The tool will allow users to explore existing HJA data sets or upload and explore their own data sets.

IV. CONCLUSIONS

We have presented the design and implementation of the Diversity Map, an interactive visualization tool and its application to the moth data set. Collaboration between ecologists, information managers, and computer scientists can potentially provide powerful tools for ecologists and managers for identifying important ecological patterns and trends as well as data sharing. We anticipate that other LTER research projects and data sets will also benefit from this kind of interactive visualization tool and collaboration.

ACKNOWLEDGMENT

Ecologists (Steven Highland, Jeff Miller, and Julia Jones), information managers (Donald Henshaw), and computer scientists (Tuan Pham and Ronald Metoyer) collaborated in this project. Funding was provided by the HJ Andrews LTER (NSF 0823380, 0218088, and 9632921), the Ecosystem Informatics IGERT (NSF 0333257), and NSF IIS-0546881.

REFERENCES

[1] P.C. Hammond and J.C. Miller. “Comparison of the biodiversity of Lepidoptera within three forested ecosystems” Annals of the Entomological Society of America 91: 323-328, 1998.

[2] S. Raimondo, A.M. Liebhold, J.S. Strazanac, and L. Butler, “Population synchrony within and among Lepidoptera species in relation to weather, phylogeny, and larval phenology” Ecological Entomology 29: 96-105, 2004.

[3] J.C. Miller, P.C. Hammond, and D.N.R. Ross, “Distribution and functional roles of rare and uncommon moths (lepidoptera: noctuidae: plusiinae) across a coniferous forest landscape” Annals of the Entomological Society of America 96(6):847-855, 2003.

[4] J. Elith and J.R. Leathwick, "Species Distribution Models: Ecological Explanation and Prediction Across Space and Time," Annual Review of Ecology Evolution and Systematics, vol. 40, 2009, pp. 677-697.

[5] C. Shannon and W. Weaver. “The mathematical theory of information,” Urbana: University of Illinois Press, 97, 1949.

[6] R. Whittaker, “Dominance and Diversity in Land Plant Communities: Numerical relations of species express the importance of competition in community function and evolution,” Science, 147(3655):250, 1965.

[7] Ed H. Chi, “A Framework for Information Visualization Spreadsheets,” Ph.D. Thesis, University of Minnesota, March, 1999.

[8] S.K. Card, J.D. Mackinlay, and B. Shneiderman, Readings in Information Visualization: Using Vision to Think, Morgan Kaufmann, 1999.

[9] T. Pham, R. Hess, C. Ju, E. Zhang, and R. Metoyer, “Visualization of diversity in large multivariate data sets,” IEEE Transactions on Visualization and Computer Graphics, vol. 16, 2010, pp. 1053-1062.

[10] D. Schuler and A. Namioka, Participatory Design: Principles and Practices, Routledge, 1993.

[11] J. Preece, Y. Rogers, and H. Sharp, Interaction Design: Beyond Human-Computer Interaction, Wiley, 2007.

[12] J.C. Miller and P.C. Hammond, “Lepidoptera of the Pacific Northwest: Caterpillars and Adults,” Forest Health Technology Enterprise Team, USDA Forest Service: Morgantown, West Virginia, 2003.

[13] A. Inselberg and B. Dimsdale, “Parallel coordinates: a tool for visualizing multi-dimensional geometry,” Proceedings IEEE Conference on Visualization, IEEE Computer Society Press, 1990, pp. 361-378.

[14] A.E. Magurran, Measuring biological diversity, Blackwell Publishing, 2004.

[15] D.L. MacAdam, “Visual Sensitivities to Color Differences in Daylight,” Journal of the Optical Society of America, vol. 32, 1942, pp. 247-274.

[16] C. Ware, Information Visualization: Perception for Design, Morgan Kaufmann, 2004.

[17] D.L. Henshaw, G. Spycher, “Evolution of ecological metadata structures at the HJ Andrews Experimental Forest Long-Term Ecological Research (LTER) site,” North American science symposium: toward a unified framework for inventorying and monitoring forest ecosystem resources, 1998, pp. 2-6

[18] S.A. Highland, “The historic and contemporary ecology of western Cascade meadows: archeology, vegetation, and macromoth ecology,” Ph.D. Dissertation, Oregon State University, Corvallis, 2011.

[19] B. Shneiderman, G. Fischer, M. Czerwinski, M. Resnick, B. Myers, L. Candy, E. Edmonds, M. Eisenberg, E. Giaccardi, T. Hewett, P. Jennings, B. Kules, K. Nakakoji, J. Nunamaker, R. Pausch, T. Selker, E. Sylvan, and M. Terry, “Creativity Support Tools: Report From a U.S. National Science Foundation Sponsored Workshop,” International Journal of Human-Computer Interaction, vol. 20, 2006, pp. 61-77.

Web Services in the U.S. Long-Term Ecological Research Network: Now and in the Future

John H. Porter[1], and Mason Kortz[2]
[1] University of Virginia
[2] University of California, San Diego
jporter@virginia.edu, mkortz@ucsd.edu

Abstract— **The U.S. Long-Term Ecological Research Network is using web services to help link data and technologies across a diverse array of ecological research sites. We review existing services that manage a dictionary of scientific measurement units and create statistical programs, and discuss future opportunities and plans for using web services to increase the effectiveness and efficiency of information management.**

Keywords— Web Services, Long-Term Ecological Research, REST

I. INTRODUCTION

The U.S. Long-Term Ecological Research Network is a collection of 26 individual Long-Term Ecological Research (LTER) sites, which engage in a wide variety of ecological research activities, and a LTER Network Office (LNO), which helps coordinate interactions between the sites and maintains network-wide databases [1]. One of the challenges faced by a network that is both so widely-distributed, and so diverse is how to provide necessary data services in a way that builds on the strengths of the sites, while minimizing duplication of effort and increasing efficiency. Web services, in the context of a service-oriented architecture, are one way to achieve the goal of efficiency, while still accommodating necessary diversity. Here we present a brief description of the LTER Network, past efforts of coordinated development activities for information management and how web services are now being used by the LTER Network. We conclude by discussing future opportunities for the development of web services.

Each of the 26 LTER sites provides funding and personnel to support information management activities at the individual site. The LTER projects tend to vary widely in their organization. Some are highly centralized, with only a small number of investigators, often at a single institution. Others are highly distributed, with large numbers of investigators spread out over a wide array of institutions. Similarly, the forms of data collected, although primarily ecologically-oriented (some LTER sites also have a special mandate for conducting social science research), also vary widely. Some sites focus primarily on long-term measurements of a relatively small number of parameters measured at a large number of locations throughout the year. Other sites measure a much wider variety of parameters, but at fewer locations or with less frequency, while still others (particularly polar and marine sites) have intense field campaigns where a large amount of data are collected in a relatively brief time period. In each case the data collection pattern represents a compromise between the scientific objectives and logistical and financial constraints. Diverse organizational, institutional and data environments for information management create variability in the structure of information management activities at the sites and the technologies they employ.

Information management at each of the sites is conducted by a staff of between one and three full-time equivalents [1]. Given the diverse array of tasks that need to be addressed by this limited staff, including preparation and management of metadata, quality assurance and control analyses, database management, and construction and maintenance of web pages, LTER Information Managers tend to be generalists, with skills using a wide array of software tools, rather than specialists in any one particular tool. Nonetheless, there are clear pockets of expertise in databases (e.g., MySQL, SQL Server, Oracle, PostgreSQL, eXist), languages (e.g., Ruby, JAVA, PHP and Perl), statistical and analytical software (e.g. R, SAS, MATLAB, SPSS) and scientific workflows (e.g., Kepler) represented in the network, albeit by information managers at different sites. Web services provide an ideal way to communicate needed information across the LTER Network because they encapsulate functionality, providing consistent machine-interpretable products, regardless of the underlying technologies used to generate that information.

II. METHODS AND TECHNIQUES

Web services offer a much-improved alternative to "screen scraping" wherein software attempts to extract needed information off a published web page [2]. Instead of the ad hoc layout of web pages, web services use structured requests to elicit structured responses across the network that are ideal for communicating information program-to-program. Both requests and responses use the well-established HTTP protocol for exchanging information. The content of the HyperText Transfer Protocol (HTTP) messages are typically serialized in eXtensible Markup Language (XML), although other serializations such as JavaScript Object Notation (JSON), Resource Description Framework (RDF), or plain text are possible.

There are at least three major ways of implementing web services: RPC (Remote Procedure Call) services, messaging services, or REST (Representational State Transfer) services

[3]. RPC and messaging services commonly use SOAP (Simple Object Access Protocol) to accept operational calls and return the results of performing those operations. This model is roughly analogous to the use of functions or subroutines in a programming language. REST services use Uniform Resource Locators (URLs) to identify and return representations of resources, rather than the results of operations [4]. REST services use the same architecture as the World Wide Web, except that the resources being accessed are machine-readable service endpoints, rather than human-readable web pages.

For the LTER Network, the Web Services Working Group (WSWG) concluded that REST-based approaches were most appropriate for the relatively simple web services required. The REST architecture is less rigid than the RPC or messaging architectures. Developers have more freedom to use exchange formats such as Ecological Metadata Language (EML) and Scientific-Technical-Medical Markup-Language (STMML) that have already been adopted by the LTER network, because the REST architecture does not require a specific exchange format, such as SOAP. The WSWG maintains a set of recommendations for sites implementing REST services to encourage a level of commonality between services distributed throughout the network [5].

III. Results and Discussion

A. Unit Registry

One of the first applications of web services in the LTER network was the management of a network-wide library of scientific units. The metadata describing data sets is most useful if the units (e.g., meters, feet) used to describe data are consistently applied. For physical measurements SI units can be applied [6], but unfortunately, there are no widely-accepted standards for describing units for many kinds of environmental data. To aid in the development of standards for describing environmental data, the LTER Network decided to create a "library" of units that would allow individuals and sites to rapidly discover what scientific units and unit descriptions were already in use and to add new units for use by others. Access to the library needed to be provided in a way that allowed a centralized, authoritative list of units to be incorporated into a wide variety of programs, web forms, and data systems. Additionally, it needed to be able to create products, such as STMML [7], for metadata construction.

The Unit Registry web service, a REST web service developed by the LTER Unit Working Group, provides read and write access to such a library of units. The web service interface supports endpoints for searching for and viewing units through the GET method, as well as creating and updating units through the POST and PUT methods. Multiple return types are supported, including XML, JSON, and plain text. A second-tier web service, the Unit Format service, was developed on top of the Unit Registry service to provide aggregate formats such as comma-separated-value (CSV) and STTML unit lists.

The Unit Registry design process began with the URL syntax (see Table 1) and the XML and JSON exchange formats. Both the web service and the clients were developed to these specifications; clients are agnostic to the specific implementation of the web service, and vice versa. Refactoring of the web service continues without interruption to the clients, provided the URL syntax and exchange formats remain unchanged. This well-documented interface allows services and clients to be developed simultaneously in a distributed environment.

The Unit Registry and Unit Format web services launched in June 2010, along with a web-based graphical user interface built on top of the service for searching and managing units (Table 1) [8]. The services are hosted by the LTER network office, developed and maintained by the Unit Working Group, and used by the entire LTER community. In the past year, 21 sites have contributed units to the Registry, and 7 have developed site tools that access the Registry via the service. These tools allow sites to draw from the shared list of units to populate local and network metadata databases. Current LTER Unit Working Group efforts are focused on improving the usability of units by creating and applying community standards for names, abbreviations, and unit-to-unit conversions. The Unit Registry is playing an important role in this process by providing an arena for information managers at all sites to collaborate. As both the content and software of the

TABLE I. Sample Unit Registry and Unit Format web services

Web Service Call	Purpose:
http://unit.lternet.edu/services/unitregistry/unit	Returns a complete list of units in the registry
http://unit.lternet.edu/services/unitregistry/unit/name=meter	Returns the unit named "meter"
http://unit.lternet.edu//services/unitregistry/unit/name~meter	Returns all units whose name contains the string "meter"
http://unit.lternet.edu//services/unitformat/stmml/unit/name=meter	Returns STMML[7] for units whose name is "meter"
http://unit.lternet.edu/services/unitformat/csv/unit/name=meter	Returns a comma-separated-value string for units named "meter"

Unit Registry continue to be developed, new tools such as unit conversions, automatic updating of deprecated units, and unit-aware workflow processes may be incorporated into the LTER network information system.

B. *Statistical Programming Service*

Another opportunity for application of web services was the creation of statistical programs for use in analyzing LTER data. The EML metadata used for documenting LTER datasets includes all of the information needed for users to construct statistical programs capable of reading and performing simple analyses (e.g., statistical summaries). A REST-based web service "statprog" fetches the requested metadata from an archive based on a unique identifier and uses XML stylesheets to transform the metadata into a statistical program (Figure 1). The statistical program that is returned can be either run, or returned to a researcher for viewing and additional editing. Automated checks made possible by the service can also be used for metadata quality control [9].

C. *Terminology Services*

In addition to the web services produced by LTER information managers, there is a wide array of web services produced by others that can be used for LTER information management. For example, the TemaTres online thesaurus software, the National Biological Information Infrastructure (NBII) Thesaurus and the Helping Interdisciplinary Vocabulary Engineering (HIVE) software provide web services that are used by the LTER Controlled Vocabulary Working Group and the LTER Network Office to augment keyword searches for data sets. Using web services, search terms are automatically expanded to include synonyms (e.g., CO_2 for Carbon Dioxide). Similarly the Integrated Taxonomic Information System (ITIS) provides web services that can be used to help automate creation of taxonomic metadata and to resolve taxonomic naming issues in data sets.

D. *Web Services and Interoperability*

As you may have noted in the previous examples, we have not specified details regarding which programming languages, databases and software tools have been used in implementing the web services. This omission is not accidental. It points out one of the great advantages of web services for a network of researchers that have expertise in different technologies: you don't need to know or understand the technology underlying a web service in order to use it.

For the REST-based services discussed here, it is sufficient to know the structure of the URL needed to invoke the web service. The product returned by the web service is similarly agnostic with respect to the tools used to process the product on the receiving end. For example, a web service call might generate an SQL query for a PostgreSQL database, that is then formatted into an XML document and returned to the requestor. That XML document then might be transformed using a stylesheet to produce a web page, analyzed using a statistical program, or be ingested into another database. The user of the web service doesn't need to know SQL in order to make the request. Similarly, the provider of the web service doesn't need to know any of the details about how the user will process the web service product. Both the web service provider and the user are able to use the tools they are most comfortable with, while effectively and efficiently transferring data using web services.

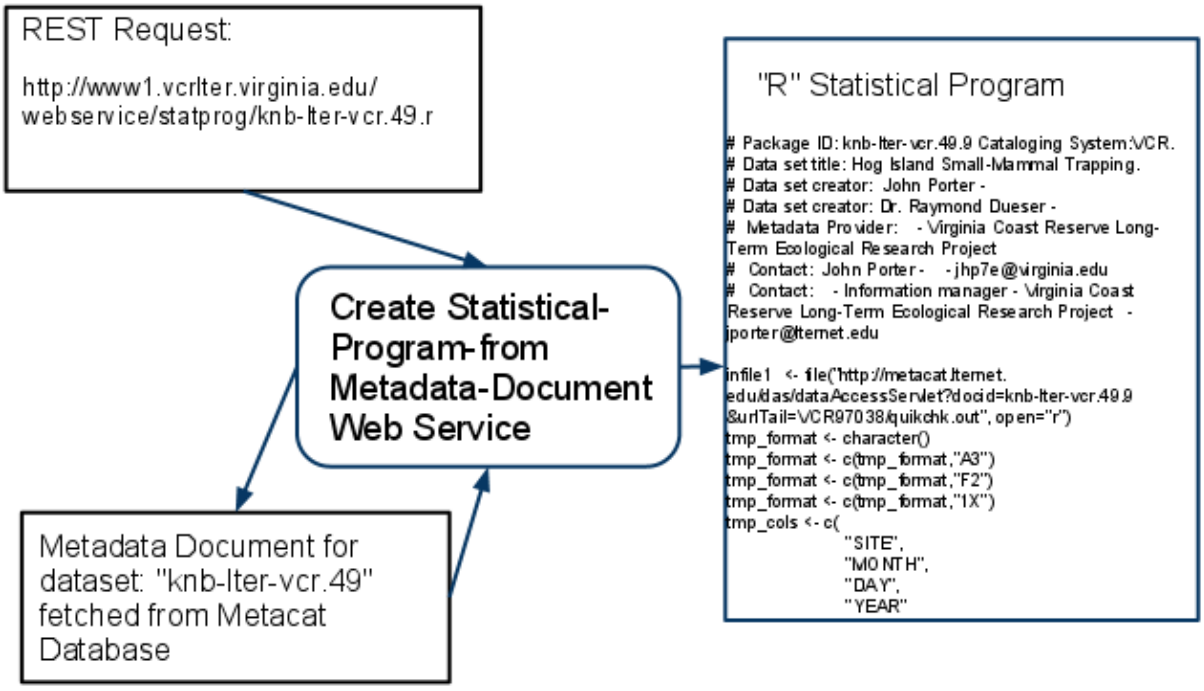


Figure 1. A REST-based web service requests a metadata document from a metadata database and transforms it into a "R" statistical program. Similar services are also available for the SAS and SPSS statistical packages just by changing the ".r" to ".sas" or ".sps" in the request.

E. Standards for Web Services

The generality of web services are facilitated by the use of community standards regarding how to structure the information exchanged through web services. For example the Simple Knowledge Organization System (SKOS) standard established by the World Wide Web Consortium (W3C) provides specifications for how lexical data used for controlled vocabularies, taxonomys and thesauri should be structured within an XML document[10]. Similarly, the Open GIS Consortium provides a variety of standards for exchange of geographical data, such as the Web Map Service specification, which specifies how maps should be shared[11]. The advantage of employing such standards in the creation of new web services is that they allow services to be widely used by a variety of clients.

F. Future Opportunities

In September 2009, the LTER Web Services Working Group (WSWG) was formed to explore the possibility of using web services in information management and make recommendations to the network for pursuing these possibilities. In addition to making general recommendations for use and development of web services within the LTER network, the WSWG was tasked with identifying specific elements within the LTER network information system that could be improved using web services.

The first network information system element the WSWG addressed was the network personnel database. This database contains roles, site affiliations, and contact information for all members of the LTER network. However, there is no mechanism in place for machine-to-machine access of this data. In February 2011, the WSWG began the process of updating the personnel database to support a REST web service interface. This interface will allow read and write programmatic access to the contents of the database. The new service, dubbed PersonnelDB, is current under development by a subcommittee of the WSWG. When PersonnelDB is completed, LTER sites will be able to seamlessly integrate information from the personnel database into web pages, and applications for creating metadata, and to provide ways of updating the database that are best suited to the organization of the individual LTER site.

The WSWG has identified other elements of the network information system that could benefit from a web service architecture. In general, any information resource that benefits from being shared among the sites from a central authoritative source, but also requires machine-level interfaces for implementation in distributed systems, is a candidate. Specific elements being looked at by the working group include bibliographies, research collaboration lists, and an expertise database. Some of these services will leverage existing web services such as the PersonnelDB service, creating a 'web of web services' or Web Oriented Architecture [12].

Outside of the WSWG, many sites are beginning to use web services, both as consumers and providers. The shift towards online, machine-readable information supports the centralization of information, and thereby a reduction in the duplication of effort throughout the network, while also supporting the decentralized use of information by allowing each site to develop their own service-enabled systems. The services can also be used by other members of the environmental research community, outside LTER. As more data, metadata, and organizational resources become available through web services, these services will become increasingly interlinked and cross-referenced. Web services that validate and enhance metadata, integrate bibliographic resources, perform quality assurance processing, and provide access to GIS data and other visualization products are all possible. In the next ten years, we hope to see the formation of a network of widely distributed, highly related, machine readable information resources - a "LTER-Wide Web" for automated agents.

G. Impacts Outside the LTER Network

Although we have largely focused here specifically on the LTER Network, web services are a nearly ideal medium for sharing capabilities and expertise with the larger community as well. Some services, such as the statistical program service, can have immediate application for any individual or group using EML as their metadata standard. Similarly, the Unit Registry holds promise for individuals and organizations interested in adopting common definitions for measurement units. As more web services come online, we expect that there will be many additional opportunities for researchers and organizations to exploit web services to expand services available and reduce duplication of effort throughout the entire ecological community.

H. Training Needs

For development and deployment of web services within the LTER Network to be completely successful, some additional training is likely to be necessary. Although most LTER Information Managers are conversant with XML, as a result of the need to create Ecological Metadata Language documents, most are less familiar with the increasingly wide variety of powerful tools and frameworks that are available for providing and using web services. Fortunately most of the applications planned in the immediate future use relatively simple XML schemas, making them easy to parse and manipulate with relatively simple tools. However, as more complex web services are deployed, familiarity with more advanced tools will be needed.

One approach that has been taken by the WSWG is to recommend that when new web services are developed within the LTER network they should be accompanied by sample web service clients that can then be modified and enhanced to meet site-specific needs. For example, deployment of the Unit Registry web service included a query interface (http://unit.lternet.edu/unitregistry) that provides both immediate utility to anyone with a web browser, and provides model code that could be modified by researchers at a particular site to meet specific needs. Providing these "models" for web service clients is a useful adjunct for informal training.

IV. CONCLUSIONS

Web services are increasingly being employed by information managers within the LTER Network to help knit together heterogeneous systems. Web services promote the sharing of information by avoiding the synchronization issues surrounding duplicated data sources, while also avoiding the security and access issues associated with providing remote access directly to databases. The use of web services allows developers at LTER sites, working groups, or the LTER Network Office to develop applications that can then be integrated into a wide variety of software systems. Consistent data and metadata, shared across many sites, in turn promotes the LTER-wide goal of scientific data integration.

In a network as diverse as the LTER, the characteristics of web services that allow them to side-step many of the traditional impediments to joint development, such as use of different types of software, languages or approaches, are especially important. For example, with web services a database expert at one site can develop a web service providing access to data in a database. That web service can then be used by an information manager at another site with expertise in analytical workflows to develop web service-accessible integrated data products that are then used by an expert in geographical information systems to produce web service-accessible maps displaying the integrated data. Use of standards and services developed outside the LTER Network will enhance the generality and power. We therefore confidently anticipate a proliferation of web services helping to meet both current and future needs.

ACKNOWLEDGMENT

Support was provided by NSF Grants 06-21014, 10-26607, and 08-23101. The LTER Web Services Working Group played a major role in the design and development of the web services discussed here.

REFERENCES

[1] Michener, W.K., Porter,J., Servilla, M. and Vanderbilt, K. 2011. Long term ecological research and information management. Ecological Informatics. 6(1):13-24. http://dx.doi.org/10.1016/j.ecoinf.2010.11.005

[2] Stein, L. 2002. Creating a bioinformatics nation. Nature 417:119-120. doi:10.1038/417119a

[3] Pautasso, C., Zimmermann, O. and Leymann, F. 2008. Restful web services vs. "big"' web services: making the right architectural decision. WWW '08 Proceeding of the 17th international conference on World Wide Web. ACM, New York, USA. ISBN: 978-1-60558-085-2, doi:10.1145/1367497.1367606

[4] Fielding, R. T. 2000. Architectural Styles and the Design of Network-based Software Architectures (Doctoral dissertation). Retrieved from http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm

[5] LTER Web Services Working Group. 2010. LTER Web Services Recommendations. Retrieved from: http://im.lternet.edu/news/committees/working_groups/webservices/recommendations

[6] National Institute of Standards. Physical Reference Data. 2003. NIST Physical Measurement Laboratory. Retrieved from: http://www.physics.nist.gov/cuu/index.html

[7] Murray-Rust,P. and Rzepa,H.S. 2002. STMML. A markup language for scientific, technical and medical publishing. Data Science, 1, 1–65.

[8] Kortz, M. 2010. Enactment and the Unit Registry. Retrieved from: http://databits.lternet.edu/fall-2010/enactment-and-unit-registry

[9] Lin, C.C., Porter J.H., Lu, S.S., Jeng, M.R. and Hsiao, C.W. (2008) Using Structured Metadata to Manage Forest Research Information: A New Approach. Taiwan Journal of Forest Science, 23, 133-143

[10] Miles, A., and Bechhofer, S. 2009. SKOS Simple Knowledge Organization System Reference. Retrieved from: http://www.w3.org/TR/skos-reference/

[11] Open Geographical Information Systems Consortium. 2006. OpenGIS Web Map Service (WMS) Implementation Specification. Retrieved from: http://www.opengeospatial.org/standards/wms

[12] Hinchcliffe, D. 2008. What Is WOA? It's The Future of Service-Oriented Architecture (SOA). Retrieved from http://hinchcliffe.org/archive/2008/02/27/16617.aspx

Toward species interaction networks – Managing, visualizing and synthesizing Gulf of Mexico geo-spatial trophic data

James Simons[1], May Yuan[2], Cristina Carollo[3], Cristina Mazza[4], Sara Gonzalez-Perez[2], Lesley Williams[2], Derek Morris[2], Dave Reed[4], Maru Vega Cendejas[5]

[1]Center for Coastal Studies, Texas A&M University-Corpus Christi
[2]Center for Spatial Analysis, Oklahoma University
[3]Harte Research Institute, Texas A&M University-Corpus Christi
[4]Fish and Wildlife Research Institute, Florida Fish and Wildlife conservation Commission
[5]El Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional

James.simons@tamucc.edu, myuan@ou.edu, Cristina.Carollo@tamucc.edu, cristina.mazza@MyFWC.com, gops@ou.edu, Lesley.N.Williams-1@ou.edu, soonernation1@hotmail.com, Dave.Reed@MyFWC.com, maruvega@mda.cinvestav.mx

***Abstract*--The last 20 years has witnessed the collection, documentation, and storage of massive quantities of biodiversity data. However, interactive networks between species that define and characterize the world's ecosystems have largely been ignored. Continued development of ecoinformatics is critical to species interaction research. Many interactions exist among species including predator-prey, competition, host-parasite, symbiosis and others. Knowledge of these interactions within an ecosystem context allows us to predict the consequences of changes in biodiversity, e.g. trophic cascades. We report on progress toward development of a model database of one type of interaction – predator/prey. Our model ecosystem is the Gulf of Mexico. Trophic data for the Gulf will be extracted from published and unpublished sources and contributed databases. Metadata "lite" has been collected for ~720 trophic references, ~650 have been geocoded and habitat information has been digitized for ~420 references. We anticipate using this network of species interactions in the context of a dynamics geographic information system (GIS) to link biodiversity data collections together in time and space. In addition, creation of an ecosystem based trophic database will have applications toward further development of food web theory, ecosystem-based fisheries models, and directed network research.**

Keywords—trophic, food web, Gulf of Mexico, metadata, CMECS, database, ecoinformatics, Gulf GAME

I. INTRODUCTION

Collection, documentation and storage of massive quantities of biodiversity data, including archiving of museum specimens and biodiversity data has evolved and amplified over the past 20 years, yet species interaction networks have largely been ignored [1]. To date, species interaction networks have been studied with small databases in the context of a very low taxonomic, spatial and temporal resolution [2, 3]. However, as larger databases are generated, ecoinformatics research will be critical to advancing the understanding of interactions among species and between species and the environment.

Museum specimen databases (e.g. FishNet2 [4], Ornithological Information System (ORNIS) [5] and others), global biodiversity databases (e.g. FishBase [6], SealifeBase [7], Encyclopedia of Life (EOL) [8] and others), and projects facilitating the use of biodiversity data (e.g Knowledge Network for Biocomplexity (KNB) [9], Global Biodiversity Information Facility (GBIF) [10], Census of Marine Life (CoML) [11] and others) provide a limited amount of species interaction data. The Interaction Web Database [12] and Webs on the Web [13] are species interaction databases for select ecosystems with only presence/absence data for that does not include interaction strength, habitat, environmental, spatial or temporal data. NOAA's Food Web Dynamics Program (FWDP) at Woods Hole, MA [14] and Resource Ecology and Ecosystem Modeling (REEM) in Seattle, WA [15], who's missions include collection, analysis and modeling of trophic interaction data, each have large collections of food habits data on slightly more than 100, mostly commercial, fish species.

Ecoinformatics emphasizes conceptual and practical tools for the understanding, generation, processing and dissemination of ecological data and information [16]. High performance computing, biologically inspired computation, object oriented data, and the internet frame informatics for ecological modeling to integrate climate, environmental, community, phenotypic and genomic data [17, 18]. Ecoinformatics explicitly recognizes the heterogeneous nature of ecological data and seeks to develop tools that consider simultaneously the high resolution and heterogeneity of the data and create added value to large volumes of data at multiple biological levels and spatial scales. Informatics research has resulted in the development of BioGeomancer [19], Lifemapper [20],

Aquamaps [21], Webs on the Web (WoW) [13], Interaction Web Database [12] and Ocean Biodiversity Informatics (OBI) [22].

Advances in ecoinformatics depend fundamentally upon database architectures that can represent entities involved in a system and the system structure across multiple taxonomic, spatial, and temporal resolutions. Key challenges posed by trophic dynamics data provide excellent ecological cases for database architecture development. The proposed research will build a trophic database for the Gulf of Mexico (GoM) to support theoretical advances in trophic dynamics. Despite the fact that many data are collected at a high level of spatio-temporal resolution (i.e., individual or size class level in each specific habitat) food web studies are not detailed, and most theory has been developed at species level (or higher) in homogeneous environments [2, 3]. This has inhibited the development of unified datasets and tools to aid development and testing of flexible, first principle, individual-based models able to explore consequences of individual variability and spatio-temporal heterogeneity of raw data which will advance the understanding of ecosystems.

II. Development of the Proposed Database

A. Database Architecture and Development

A spatio-temporal database architecture for ecological interactions will be designed to account for the heterogeneity of trophic data. The complexity and diversity of the data creates challenges in building an ecoinformatics database. Because our approaches to data representation and organization will center on complex system processes and ecological interactions, as well as account for data heterogeneity, the database architectures developed will be transferable to other ecological domains.

We will adopt Hierarchy Theory to develop database architectures that address common ecological issues, such as grain and scale, identification of entities, levels of dynamics, and disturbances [23]. Hierarchy by definition imposes ordinations, as from smaller to larger, or from simpler to more complex. These concepts from Hierarchy Theory are central to many complex systems, including ecological systems and weather systems [24]. Database architectures built upon these concepts will provide rich grounds for data mining and knowledge discovery of higher level concepts [25].

Database architecture includes two components: (1) representation of reality; and (2) organization of data. The first component concerns what concepts or objects need to be represented in the database and how to most effectively represent these concepts or objects in database models. Because our proposed research aims to integrate spatial and temporal information for ecological interactions, we need to represent spatial and temporal characteristics of the identified concepts or objects. The second component addresses how different sets of data, such as species, habitat, sea surface temperature, management zones, *etc*, should be organized in the ecoinformatics to support modeling efforts that relate multiple variables to derive new understanding or forecasting. Both components of data representation and data organization need to account for complexity and diversity of ecological systems and the nature of potential data sources.

B. Data Sources, Acquisition, and Quality Assurance

This project will encompass the marine and estuarine waters of the GoM, along the United States, Mexico and Cuba. Species that inhabit the Gulf region and its waters for at least part of their life cycle will be included, *eg*. taxonomic groups listed in Table 1. Habitats covered include estuaries and continental shelf as well as the pelagic, mesopelagic, continental slope, and abyssal realms.

TABLE 1, TAXONOMIC GROUPS TO BE INCLUDED IN THE GOM TROPHIC DATABASE AND THE CURRENT STATUS OF IN-HAND AND PERCEIVED REFERENCES ADDRESSING FOOD HABITS.

Taxonomic Group	**Number of References in Hand**	**Estimated References Available**	**Total Species Currently Cited with Diet Data**
Marine Mammals	3	25	1
Sea Turtles	9	10-15	3
Fishes	721	740	~650
Sea and Shore Birds	4	100-200	4
Crustaceans	19	25-50	58
Mollusks	3	25	45
Polychaetes	~25	100-200	99
Ctenophores	5	10	2
Cnidarians	5	10	6

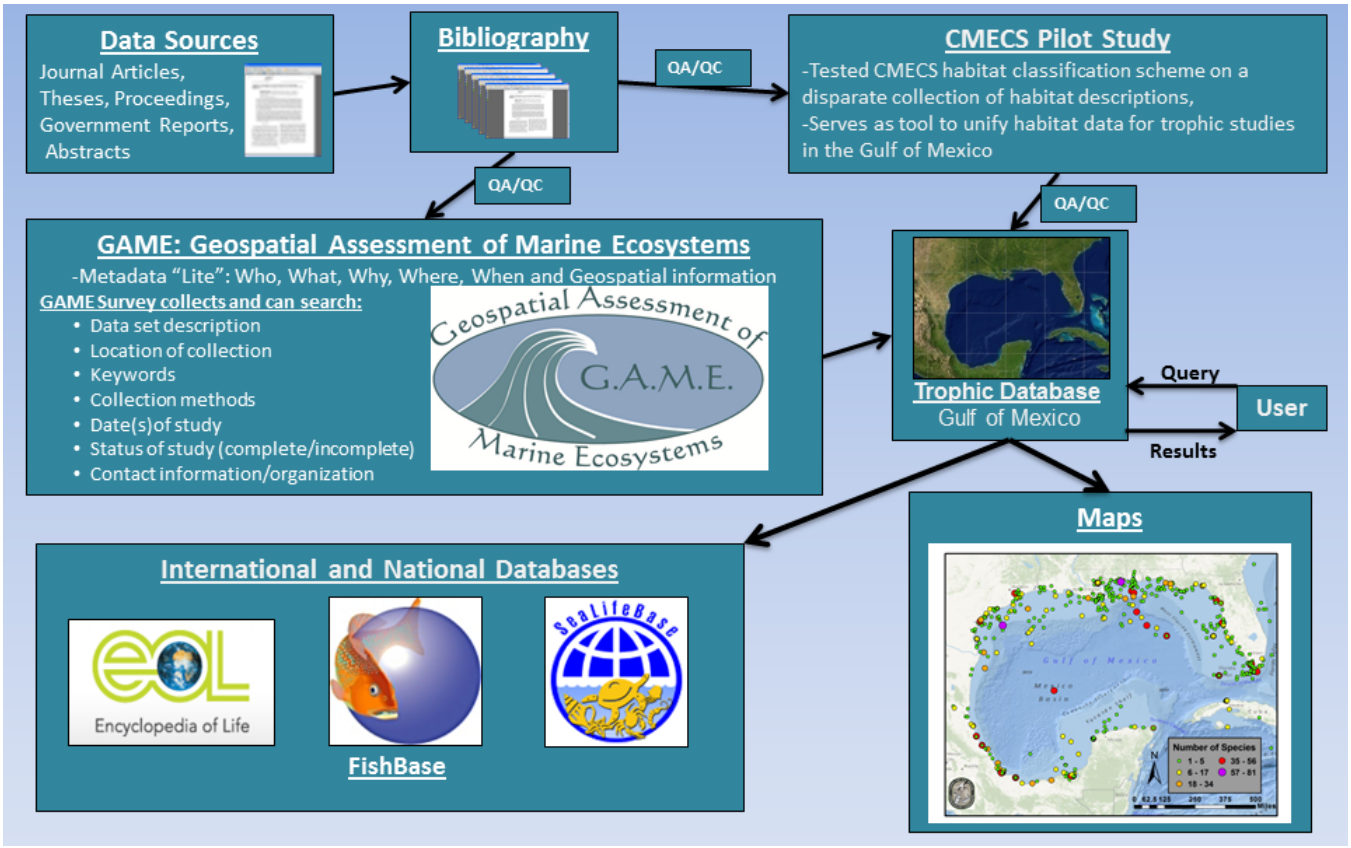


Figure 1. Schematic of the GoM trophic database workflows, links and outputs.

The following categories of data will be extracted from each source, when provided: Geopolitical location, Geospatial areas, Habitat, Geographic location, Time, Physico-chemical data, Collection method, Taxonomy, Specimen data, Food description, Stable isotopes and Source. Draft metadata fields as well as data and function requirements analysis will be developed. The database schema will follow the Ecological Metadata Language (EML) [26], an Extensible Markup Language (XML)-based metadata specification, and OBIS schema to ensure we structure marine data properly (Fig. 1). As part of this process, we will contribute metadata standards for trophically related data.

Data will be extracted from peer reviewed articles, government reports, dissertations/theses, abstracts, conference proceedings, electronic databases and unpublished data. Our data entry system will have error checking routines built into a data entry interface. Data available in electronic document will be extracted with wrappers. When feasible, tabular numeric hard copy data will be scanned with optical character recognition (OCR) software and converted to an electronic format for manipulation and extraction. Graphical data will be scanned into digital format. Data quality will, to some extent, be maintained through users reporting errors, similar to The Paleobiology Database [27] and other community-based cyber-infrastructure. Spatial context of the data will be preserved, through maps, names, coordinates or descriptions of sampling locations. Spatial data will be documented with the Federal Geographic Data Committee (FGDC) Biological Profile [28] and metadata made available with the FGDC Clearinghouse mechanism. Metadata will provide the user adequate information to make an assessment of the quality to ensure informed use of the data.

C. *Informatics Tools*

To access, process and create value-added analyses, informatics tools will be developed or links provided to websites with existing tools. We will create an interactive, spatial analyst tool for accessing, analyzing, visualizing, and production of distribution maps of predator and prey and other spatially based graphic displays of diet data. Users will select and access physico-chemical, habitat, geo-political, or other variables relevant to the study of predator-prey relationships. Temporal data will be used to evaluate the effects of environmental and climate change on trophic dynamics and evolutionary processes. In addition, these data will be useful for: assessing bioaccumulation and trophic transfer of historic and newly emerging contaminants [29], joining large biodiversity datasets together for better trophic ecosystem models [30] and drawing various inferences on the ecological functioning and fisheries impacts [31].

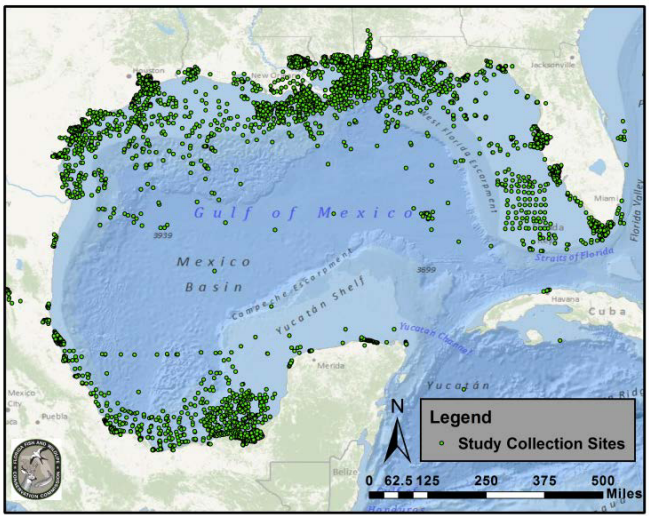


Figure 2. Map showing the location of individual sampling sites for ~520 food habits studies in the GoM.

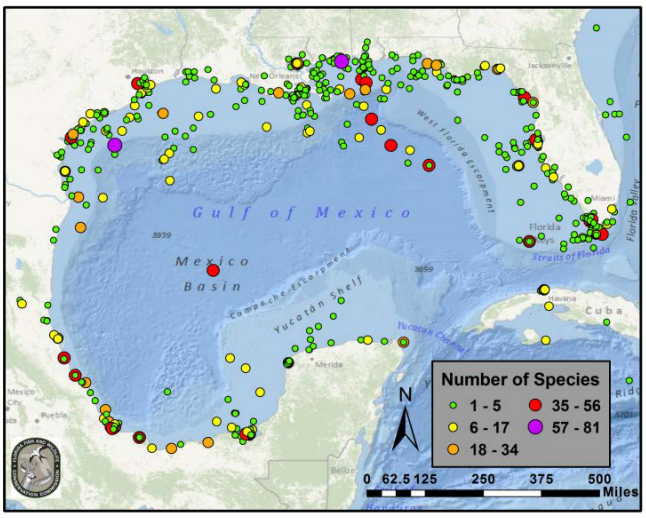


Figure 3. Map showing the centroid location of ~520 food habit studies and the number of fish species examined for food habits.

A metaweb, using the raw data without any a priori aggregation, will be flexible, (i.e., individual based to species level, homogeneous to heterogeneous space, etc.) to explore consequences of individual variability, and spatio-temporal heterogeneity of the raw data, and level of taxonomic, spatial and temporal aggregation for understanding of ecosystems. Fuzzy kriging techniques [32] will incorporate both crisp (certain) and fuzzy data to estimate categorical regions (such as abundance or average) of species distributions or trophic relations. Self-organizing maps (SOM) [33] will be developed to measure similarity of trophic structures in different habitats. A SOM will show clusters of habitats based on their trophic characteristics. Other informatics tools include qualitative reasoning models for trophic interactions among populations [34], genetic algorithms to predict food habits of fishes in unstudied habitats [35] and adaptive agents to simulate food webs [36].

D. Web Applications

Data will be publically available through a multi-lingual website with relational database and geographic information system (GIS) entry portals. Data will be available on the website in two formats: 1) table format; and, 2) EML formats for the purpose of information exchange with other databases. To exchange data with other databases, server software such as Distributed Generic Information Retrieval (DiGIR) or Taxonomic Database working Group (TDWG) Access Protocol for Information Retrieval (TAPIR) will be adapted to send/retrieve data on the Internet. Links will be provided to relevant database and informatics websites [5, 6, 7, 8, 9, 10, 11, 12, 13, 18, 20, 22 and others].

E. Challenges

Creating and using the proposed GoM trophic database presents several challenges. The various studies were conducted under a wide variety of objectives and methods, requiring units and methods be standardized to the extent possible. Data are reported in a wide array of graphic and tabular formats, which will need to be converted to a single database format. The spatial and taxonomic distribution of the species is clumped (Fig 2), requiring the use rarefaction and interpolation where feasible. While these issues present challenges in analyzing these data, they also identify opportunities for further research.

III. Results and Discussion

A. Geo-Coding References

We began by capturing the spatial information for the studies (i.e., station locations, and locations and names of systems where the studies were conducted) and display the results in a GIS (Fig 1). Study collection points, polygons, and centroid points (derived from the study polygons) have been created for ~650 of the ~720 references at the University of Oklahoma and the Florida Fish and Wildlife Research Institute. Attributes of these points, polygons, and centroids include the study's author, study location, number of species studied (Fig 2 and 3), and associated metadata.

B. *Coding the Coastal and Marine Ecological Classification Standard (CMECS)*

A pilot study was conducted at the University of Oklahoma to unify codification of habitat data in the numerous trophic references using CMECS [37]. Approximately 60% of the references in hand at the time the project was undertaken were coded. This entailed extracting all relevant habitat information reported in the document and adapting those descriptions to the CMECS terminology. The CMECS system first classifies a habitat into one of two systems, and then up to five components (Water Column, Benthic Biotic, Surface Geology, Sub-benthic, GeoForm) can be used to provide detailed information.

C. *Metadata and Gulf Geospatial Assessment of Marine Ecosystems (GAME)*

Metadata records were created for 690+ peer-reviewed papers organized in a Food Habits of Fishes Bibliography for estuarine and marine environments of the Gulf of Mexico [38]. Metadata were generated using the Gulf GAME survey tool that allows records to be entered through a user friendly interface. These records were incorporated into the Gulf GAME catalog and are available online for search and retrieval [39]. The catalog stores metadata "lite" (i.e. only primary elements are captured) and the records are FGDC compliant. The importance of this work lies in that it allows archival for long-term persistence of information that previously had no attendant metadata. Also, it makes the information discoverable since the majority of the Bibliography studies are not available online.

IV. CONCLUSIONS

The trophic informatics system for the GoM will be designed to be compatible and extensible to extant database projects and programs. Coordination and collaborative promises have already been achieved with FishBase [6], SeaLifeBase [7] and EOL [8]. This will allow for data and format sharing so that the maximum accessibility and usefulness of the trophic data are achieved, and that value is added to the existing databases through pre-planned links. The vision is that the structure, methods, and tools will be extensible to other large marine ecosystems. The extensibility and transportability of the model is important to support the development of similar databases globally. Toward this end, this database will prove invaluable in furthering research on directed networks, ecosystem fisheries models and food web theory.

ACKNOWLEDGMENT

This work was supported in part by NatureServe and NOAA for a CMECS pilot study and the U.S. Environmental Protection Agency Gulf of Mexico Program provided funds to create metadata "lite" records and to capture spatial information from a collection references on food habits of GoM fishes provided by Dr. J. Simons. Comments from three anonymous reviewers greatly improved the manuscript.

REFERENCES CITED

[1] K. McCann, "Protecting biostructure," Nature, vol. 446, p. 9, 2007.

[2] R. M. May, "Network structure and the biology of populations," TREE, vol. 21, number 7, pp. 394-399, 2006.

[3] T.C. Ings, J.M. Montoya, J. Bascompte, N. Blüthgen, L. Brown, C.F. Dormann, F. Edwards, D. Figueroa, U. Jacob, J.I. Jones, R.B. Laurisden, M.E. Ledger, H.M. Lewis, J. Olesen, E.J.F. van Veen, P.H. Warren, and G. Woodward, "Ecological networks -- beyond food webs," J. Anim. Ecol., vol, 78, pp. 253-269, 2009.

[4] FishNet2, 2008, http://fishnet2.net/, Accessed 8/2011.

[5] Ornithological Information System, 2008, http://olla.berkeley.edu/ornisnet/, Accessed 8/2011.

[6] Froese, R., Pauly, D. (Eds), 2009, FishBase, World Wide Web electronic publication, www.fishbase.org, version (06/2009).

[7] Palomares, M.L.D and Pauly, D. (Eds.), 2009. SeaLifeBase. World Wide Web electronic publication. http://www.sealifebase.org/, Version (3/2009).

[8] Encyclopedia of Life, 2008, http://www.eol.org/, Accessed8/2011.

[9] Knowledge Network for Biocomplexity, 2009, http://knb.ecoinformatics.org/index.jsp, Accessed 8/2011.

[10] Global Biodiversity Information Facility, 2008, http://www.gbif.org/, Accessed 8/2011.

[11] Census of Marine Life, 2008, http://www.coml.org/, Accessed 8/2011.

[12] Interaction Web Database, 2008, http://www.nceas.ucsb.edu/interactionweb/, Accessed 8/2011.

[13] Webs on the Web, 2008, http://foodwebs.org/, Accessed 7/2011.

[14] FWDP, 2011, http://www.nefsc.noaa.gov/pbio/fwdp/FWDP.htm, Accessed 8/2011.

[15] REEM, 2011, http://access.afsc.noaa.gov/reem/ecoweb/Index.cfm, Accessed 8/2011.

[16] F.A. Bisby, "The quiet revolution: biodiversity informatics and the internet," Science, vol. 289, pp. 2309-2312, 2000.

[17] W.K. Michener, J.W. Brunt, and K.L.Vanderbilt, "Ecological informatics: a long-term ecological research perspective," in Proceedings Information Systems Development II, N.J. Callaos, J. Porter, and N. Rishe, Eds, 6th World Multiconference on Systemics, Cybernetics and Informatics, 2002.

[18] M.B. Jones, M.P. Schildhauer, O.J. Reichman, and S. Bowers, "The new bioinformatics: Integrating ecological data from the gene to the biosphere," Ann. Rev. Ecol. Evol. Syst., vol. 37, pp. 519-544, 2006.

[19] BioGeoMancer, 2002, BioGeoMancer: Automated Georeferencing for Natural History Collections, http//www.biogeomancer.org, Accessed 8/2011.

[20] LifeMapper, 2008, http://www.lifemapper.org/, Accessed 8/2011.

[21] Aquamaps, 2008, http://www.fishbase.ca/tools/aquamaps/search.php, Accessed 8/2011.

[22] Ocean Biodiversity Informatics, 2008, http://www.vliz.be/events/obi/index.php, Accessed 8/2011.

[23] V. Ahl, and T.F.H Allen, Hierarchy Theory: A Vision, Vocabulary, and Epistemology. New York: Columbia University Press, 1996.

[24] M. Yuan, "Representing geographic information to enhance GIS support for complex spatiotemporal queries," Trans. in GIS, vol. 3, no. 2, pp. 137-160, 1999.

[25] M. Yuan, "Knowledge discovery of geographic dynamics in spatiotemporal data," in Geographic Data Mining and Knowledge Discovery (2nd ed), H. Miller and J. Han, Eds. CRC/Taylor and Francis. 2008.

[26] Ecological Metadata Language, 2008, http://knb.ecoinformatics.org/software/eml/, Accessed 8/2011.

[27] The Paleobiology Database, 2008, http://paleodb.org/, Accessed 7/2011.

[28] Federal Geographic Data Committee (FGDC), 2008, http://www.fgdc.gov/, Accessed 8/2011.

[29] P.A. Sandifer, A.F. Holland, T.K. Rowles, and G.I. Scott, "The oceans and human health," Environ. Health Perspect, vol. 112, no. 8, pp. 454-455, 2004.

[30] R.A. Myers, J.K. Baum, and T.D. Shepherd, "Cascading effects of the loss of apex predatory sharks from a coastal ocean," Science, vol. 315, pp. 1846-1850, 2007.

[31] L. Vidal, and D. Pauly, "Integration of subsystems models as a tool toward describing feeding interactions and fisheries impacts in a large marine ecosystem, the Gulf of Mexico," Ocean Coastal Management, vol. 47, pp. 709-725, 2004.

[32] A. Salaski, "Ecological applications of fuzzy logic," in Ecological Informatics: Scope, Techniques and Applications, 2nd ed. F. Recknagel Ed. New York, NY: Springer, 2006, pp. 3-14.

[33] J.L Giraudel, and S. Lek, "Ecological applications of non-supervised artificial neural networks," in Ecological Informatics: Scope, Techniques and Applications, 2nd ed. F. Recknagel Ed. New York, NY: Springer, 2006, pp 49-67.

[34] P. Salles, B. Bredeweg, S. Araujo, and W. Neto, "Qualitative models of interactions between populations," AI Communications, vol 16, no. 4, pp. 291-308, 2003.

[35] D.J. D'Angelo, L.M. Howard, J.L. Meyer, S.V. Gregory, and Z.L.R. Ashkenas, "Ecological uses for genetic algorithms: predicting fish distributions in complex habitats," Can. J. Fish. Aquatic Sci. vol. 52, pp. 1893-1908, 1995.

[36] F. Recknagel, "Ecological applications of adaptive agents," in Ecological Informatics: Scope, Techniques and Applications, 2nd ed. F. Recknagel Ed). New York, NY: Springer, 2006, pp 109-124.

[37] M. Yuan, L. Williams, S. Gonzalez-Perez, D. Morris, and J. Simons, Data acquisition and meta-analysis of habitat information from Gulf of Mexico trophic studies using CMECS, Final Report submitted to NatureServe. NOAA contract # EA133C-05-CQ-1051 and Fugro EarthData Project # E09-0039-00. 22p. 2010.

[38] J.D. Simons, R.M. Darnell and M.E. Vega-Cendejas, Bibliography of studies of Food Habits of Estuarine and Marine Fishes in the Gulf of Mexico, unpublished manuscript.

[39] FWRI, 2011, http://myfwc.com/research/gis/game/, Accessed 7/2011.

Lifemapper: Infrastructure and Services for Biodiversity Science

Aimee Stewart[1], James Beach[1], C. J. Grady[1], Jeffrey Cavner[1]
[1] University of Kansas, Biodiversity Institute
astewart@ku.edu, beach@ku.edu, cjgrady@ku.edu, jcavner@ku.edu

***Abstract*—Lifemapper is an archive of species and environmental data, predicted habitat maps and a suite of data and analysis web services based on these data and the computational processes used to create them. Behind the scenes, Lifemapper relies on open source software libraries, modular code design, and a collaborative development process. As a community resource, Lifemapper is committed to standard data formats and Internet access protocols and is increasingly focused on data transparency and repeatability through cataloging and documenting metadata and provenance.**

Keywords—biodiversity; geospatial; species distribution modeling; macroecology; metadata; standards; infrastructure; web services

I. INTRODUCTION

Lifemapper (www.lifemapper.org) is a computational infrastructure project funded by the National Science Foundation (NSF) that combines open source geospatial and biodiversity informatics tools to: enable biogeographical analyses of current and future distributions of species, demonstrate the biological impacts of climate change to junior and senior high school students, and increase the research utilization of the data associated with biological specimens housed in museums around the world. Lifemapper (LM) is organized around two primary components: 1) an archive of predicted current and future species distribution maps and, 2) a set of software tools and services that enable biological researchers to predict and analyze single- and multi-species, multi-scale patterns of species distribution. Lifemapper's software architecture includes a data pipeline that moves researcher requested modeling experiments to a 64-node cluster for computation, and then retrieves the results. Lifemapper then catalogs resulting model outputs, datasets, statistics and metadata for retrieval through standardized web services defined by Open Geospatial Consortium (OGC, http://www.opengeospatial.org/) standards and simple Representational State Transfer (REST) [1] architectural style.

II. ARCHIVE

The first Lifemapper component is an extensive archive of predicted species habitat maps. LM's species distribution modeling (SDM) data pipeline automatically assembles experiments with available species occurrence data and with current and future scenario climate data. The input species occurrence data used by LM are aggregated from biological museums, collections and observation databases by the Global Biodiversity Information Facility (GBIF, http://data.gbif.org/). LM calculates SDM experiments from GBIF specimen data and climate data using openModeller (http://openmodeller.sourceforge.net) [2], an open source species modeling framework, which supports a number of ecological niche modeling algorithms as plug-ins, including the most widely-used methods: GARP with Best Subsets [3], Bioclimatic Envelopes [4,5] and Maxent [6]. Climate data includes bioclimatic variables from Worldclim (http://www.worldclim.org) and Global Climate Model (GCM) outputs distributed by the UK Met Office Hadley Centre (http://www.metoffice.gov.uk/climate-change/resources/hadley/) and the National Institute for Environmental Studies, Japan based on International Panel on Climate Change (IPCC) defined scenarios for the Third Assessment Report (TAR, http://www.ipcc-data.org/gcm/monthly/SRES_TAR/index.html) and Fourth Assessment Report (AR4, http://www.ipcc-data.org/gcm/monthly/SRES_AR4/index.html). LM maintains an archive of automatically generated niche model maps, as well as the input species occurrence and climate data used in their creation, for public exploration and retrieval through the Lifemapper web site and web services.

The Lifemapper SDM Pipeline connects the data archive and the computational processes to monitor the system for user-requested experiments and updated specimen data from GBIF, which trigger initial or re- calculation of affected experiments. Worker threads simultaneously update experiment status and inputs, submit experiments to and retrieve results from a 64-node compute cluster. Once results have been written to storage, and metadata cataloged in the system, they are immediately available through LM web services.

III. WEB SERVICES

Lifemapper provides the second component, a set of geospatial data and analysis capabilities for use with the LM archive or user data, as web services. All Lifemapper web services are available as web applications at http://www.lifemapper.org, but also can be accessed programmatically using simple Uniform Resource Locator

(URL) construction to identify the web service and appropriate parameters. LM data web services serve specimen occurrences, environmental datasets and predicted habitat maps, as well as metadata for all these data layers.

A. Species Distribution Modeling

Analysis tools include Species Distribution Modeling (LmSDM) services available through a REST and OGC Web Processing Service (WPS) interfaces. LmSDM services can be requested using either user-supplied or LM-provided data, and offer model calculations using openModeller and the algorithms implemented within that framework.

As part of the Kansas-Oklahoma NSF EPSCoR project "A Cybercommons for the Great Plains" effort, Lifemapper developed plug-ins for VisTrails scientific workflow software (http://www.vistrails.org), developed by the Scientific Computing and Imaging Institute at the University of Utah, to simplify LmSDM access. This plug-in integration between LM web services and the VisTrails workflow environment enables climate change scientists to assemble complex computational pipelines consisting of sequential tasks connected through an intuitive drag-and-drop programming user interface on the desktop. The LM-VisTrails plug-in enables users to design species distribution modeling experiments using LM data and LmSDM web services to run a species distribution modeling experiment. As additional web services move to production, the LM-VisTrails plugins will include those services as well.

B. Range and Diversity

In collaboration with the University of Connecticut (R. Colwell, T. Rangel) in the NSF project "Extending Lifemapper to Enable Macroecological Research", Lifemapper: Range and Diversity (LmRAD) explores the biogeography of species and biodiversity of regions. LmRAD focuses on two fundamental units of biogeography: species range and species diversity. It creates species Presence-Absence Matrices (PAMs), an approach for linking patterns of range size and of species richness at biogeographical scales [7]. The PAM is a gridded data format, where the x-axis represents species and the y-axis represents geographic sites. Each matrix element is coded for the presence (1) or absence (0) of each of hundreds or thousands of species at a given site, by intersecting species range data layers with a grid representing the area of interest. PAMs are the starting points for multiple methods used to test ecological and evolutionary hypotheses about the spatial patterns of biological diversity on continental and global scales.

Arita et al. [8] have shown there are correlations between: a) the species diversity of site (marginal total of diversity) and the mean range size of all species within that site, and b) between the range size of a species (marginal total of occupancy) and the mean species diversity within the range of that species. The correlations are mirror images of the same pattern, reflecting fundamental mathematical and biological relationships represented by the PAM. Range-diversity scatter plots depict these relationships graphically by-species and by-site. After computing indices, the grid is randomized and the process repeated to assess the significance of results.

Lifemapper is concurrently developing plug-ins to Quantum-GIS (QGIS, http://qgis.osgeo.org), a versatile open source Geographic Information System (GIS) desktop application, to simplify access to the LmRAD modules and visualize experiment inputs and results in a full-featured GIS application. By using the multi-platform QGIS as a client to the LmRAD services, Lifemapper brings a powerful set of macroecological analysis tools to a wide variety of users, regardless of the computational power or operating system of their desktop computer. All outputs are provided in standard formats, to simplify further analysis in other software applications.

IV. Guiding Principles

A. Research and Education

Students and educators are the main focus of Lifemapper archive creation. The LM archive presents overall picture of predicted distributions for species with adequate digital data. In the NSF Education-funded collaborative project "Change Thinking for Global Science" with the University of Michigan, we are building progressive learning sets with curricula using targeted species in the LM archive to teach middle school students complex concepts of science and ecology. In these learning sets, we have created online worksheets that present material about weather, climate, species, and allow exploration of species distribution maps predicted for current day, and three time steps in the future. Online worksheets guide students through the material to build upon knowledge gained in previous exercises.

Undergraduate students, graduate students, and researchers are the intended audience for data and analysis services, and client tools created to access them. Graduate and post-graduate researchers may use the client applications to easily create a suite of experiments comparing results between different datasets, parameters, and geographic scale. As our data and metadata publishing system goes into production, the metadata available for datasets and provenance information available for experiments will allow researchers to reference and publish input data or an entire experiment with parameters and explanatory annotations referenced in a peer-reviewed publication.

B. Standards facilitate interoperability

Running through all aspects of the Lifemapper project is a commitment to using data and communication standards. LM services adhere to well-defined standards giving developers a clear framework to work within, and providing LM users a service where issues and solutions are well documented. Metadata web services are based on the REST service model.

Lifemapper implements four OGC standards. Web Processing Service (WPS) is a standard that defines an interface for publishing geospatial processing services, defines how a client may request those services, and standardizes requests and responses. Web Mapping Service (WMS), allows

simple rendering of one or more spatial datasets. Two data services, Web Feature Service (WFS) and Web Coverage Service (WCS) return XML formatted vector data and raster datasets respectively. All of these OGC services interact with geospatial data in standard formats supported by the GDAL/OGR (http://www.gdal.org) geospatial library.

C. Metadata empowers data

An important principle underlying Lifemapper data and services is that consistent metadata should be available concurrently with LM-associated data and analyses. Accurate metadata is the cornerstone of data discovery and re-use. All static LM data will be publicly cataloged and LM web services will allow users to catalog metadata for LM-generated data and experiments with varying degrees of public access. Metadata can currently be requested for any LM-generated data in Ecological Metadata Language (EML, http://knb.ecoinformatics.org/software/eml), a format ideal for a wide range of ecological datasets [9], with plans to offer other relevant formats in the near future.

To provide a more detailed description of the procedures performed in an LM experiment, Lifemapper is extending the process module of EML. This extended EML moves Lifemapper closer to the goal of creating full provenance documents containing a history for any research experiment. The LM EML Reader module then enables re-execution with the same or modified inputs and parameters to replicate or produce variations on the documented experiment. The metadata can be published with journal articles, linking the research to the inputs and software, code or web services used to perform the processing. The LM-VisTrails and LM-QGIS plugins contain the EML Reader allowing experiments to be recreated in those software applications. Lifemapper is also expanding the EML Reader to transform LM experiment metadata into narratives, suitable for different audiences. As these EML extensions are refined, Lifemapper will submit them to the EML working group to consider for inclusion in the standard.

V. Moving Forward

A. Lessons Learned

As the Lifemapper project has matured and expanded, the importance of a flexible codebase has become increasingly apparent. The Lifemapper project follows the object-oriented programming paradigm, with particular emphasis on modularity, inheritance, and data abstraction. All code is written in Python (http://www.python.org), an open-source, cross-platform, high-level language that facilitates rapid development and easy debugging.

As the project has expanded to encompass additional data and services, we have discovered areas of the code that were overly specific. As we encounter modules that are difficult to extend, we revisit the design of the module and refactor, often creating a more complex object hierarchy, or following an accepted software design pattern [10].

Similarly, after switching to a heterogeneous cluster environment, we generalized the scheduling code that distributes analysis jobs among cluster nodes. The new design enables us to distribute different types of jobs to a variety of compute engines, both local and remote.

As a team, we have increased our cohesiveness and adaptability and clarified our shared vision by adopting a modified Scrum [11] approach (http://www.scrum.org/) to Agile software development (http://agilemanifesto.org/), which emphasizes iterative and incremental software development. We use a Trac (http://trac.edgewall.org/) wiki and issue tracking system with plug-ins integrating a Subversion code repository and Agilo for trac (http://www.agilofortrac.com/), to set goals, document decisions, establish milestones, determine the tasks and subtasks required to reach those milestones, and track timelines and progress. This system has increased accountability, while giving all team members a clear vision of the road ahead.

B. Onward

As a core component of the NSF Experimental Program to Stimulate Competitive Research (EPSCoR) Cybercommons project, LM is committed to becoming a contributing node of the NSF Data Observation Network for Earth (DataONE, http://www.dataone.org). DataONE is a $20M, 10-year collaboration among several universities (including KU, UNM, Oak Ridge National Labs, and the National Center for Atmospheric Research) whose mission is to build sustainable, long-term infrastructure for storage, indexing, discovery and access to earth observation data. Data sets cataloged within the DataONE system will be available through a set of well-defined application programming interfaces (APIs) for analytical research client packages. By implementing the DataONE APIs for data and metadata, LM will connect to a community-standards based distributed repository which will archive LM-facilitated research and modeling outputs and promote wide interoperability and integration within the computational earth science community.

As part of the ChangeThinking and LmRAD grants, our vision is to expand our educational resources to target graduate researchers as well as high school students. Our website will include guided documentation explaining and documenting previous research in SDM, algorithm strengths and weaknesses, the effect of various input parameters, limiting environmental factors, macroecological indices, species attributes affecting dispersal limits, and more. References to publications relevant to Lifemapper resources will be cited and provide a primer for students new to the field.

Our next collaboration expands the environmental data we provide for LmSDM and LmRAD to include NASA Earth observational data through a partnership with University of New Mexico (UNM) Earth Data Analysis Center (EDAC) and University of Texas at El Paso (UTEP) Cyber-ShARE Center. Cyber-ShARE provides an instrumental approach for collecting provenance information, the CI-Miner Method [12], developed at University of Texas at El Paso (UTEP). This project will instrument both EDAC and LM services to

capture end-to-end provenance within and across these two platforms.

C. Conclusion

Lifemapper's contribution to the biodiversity science infrastructure began with a simple vision of computing species distribution maps for available digital specimen data. It has grown to provide analysis and data web services to middle school students, researchers, and external applications. Lifemapper will continue to expand offerings of geospatial biodiversity data, computational resources, metadata, and research documentation in standard formats through community portals and well-publicized APIs to make data and research created with Lifemapper tools more accessible, reliable, and trustworthy.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant Numbers EPS 0919443, DRL 0918590, BIO/EF 0851290, OCI 0753336. The authors are grateful for intellectual discussions with Jorge Soberón, Professor, Ecology & Evolutionary Biology, University of Kansas, Andres Lira and Narayani Barve, Graduate Students, Ecology & Evolutionary Biology, University of Kansas.

REFERENCES

[1] R.T. Fielding. "Architectural styles and the design of network-based software architectures," Doctoral dissertation, University of California, Irvine, 2000.

[2] V.P. Canhos, S. Souza, R. de Giovanni and D.A.L. Canhos. 2004. "Global biodiversity informatics: Setting the scene for a "New World" of ecological modeling," Biodiversity Informatics 1: 1-13.

[3] R. P. Anderson, D. Lew, and A. T. Peterson. "Evaluating predictive models of species' distributions: criteria for selecting optimal models," Ecological Modelling, vol. 162, pp. 211-232, 2003.

[4] H.A. Nix, "A biogeographic analysis of Australian Elapid snakes," Atlas of Elapid Snakes of Austrailia, vol. 8, R. Longmore, Ed., pp. 4-15, 1986.

[5] J. R. Busby, "BIOCLIM – A bioclimatic analysis and prediction system," Nature Conservation: Cost Effective Biological Surveys and Data Analysis, C.R. Margules and M.P. Austin, Eds., Canberra: CSIRO, 1991, pp. 64-68.

[6] S.J. Phillips, R.P. Anderson and R.E. Schapire, "Maximum entropy modeling of species geographic distributions," Ecological Modelling, vol 190, pp. 231-259, 2006.

[7] McCoy, E. D., and K. L. Heck, Jr. 1987. "Some observations on the use of taxonomic similarity in large-scale biogeography," Journal of Biogeography 14:79–87.

[8] H.T. Arita, J.A. Christen, P. Rodríguez, and J. Soberón, 2008. "Species diversity and distribution in presence-absence matrices: mathematical relationships and biological implications," The American Naturalist 172: 519-532

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

[10] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, 1995. Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley. ISBN 0-201-63361-2.

[11] K. Schwaber, M. Beedle, 2002. Agile software development with Scrum. Prentice Hall. ISBN 0130676349.

[12] P. Pinheiro da Silva, L. Salayandia, A. Gandara, A.Q. Gates, "CI-Miner: semantically enhancing scientific processes," Earth Science Informatics vol. 2, no. 4, pp. 249-269, 2009.

.

DataONE

Promoting Data Stewardship Through Best Practices

Carly Strasser[1,2], Robert Cook[1,3], William Michener[1,4], Amber Budden[1,4], Rebecca Koskela[1,4]
[1] DataONE
[2] University of California Santa Barbara
[3] Oak Ridge National Laboratory
[4] University of New Mexico

strasser@nceas.ucsb.edu, cookrb@ornl.gov, wmichene@lternet.edu, aebudden@dataone.unm.edu, rkoskela@unm.edu

***Abstract*—The ecological and environmental sciences are comprised of many different disciplines, each with their own methods, theories, and culture. A characteristic that most of these different disciplines share, however, is a lack of culture for good stewardship of data. Characteristics of good data stewardship include understanding the importance of data management, using best practices for managing data, and recognizing the value of data sharing and data reuse for the future of ecology and the environmental sciences. The Data Observation Network for Earth (DataONE) is actively developing a community database of best practices that can be easily accessed and adopted by scientists to promote good data stewardship practices and lead to high quality data products. Here we introduce DataONE's approach to developing the best practices database and provide a data management primer that contains examples relevant to all elements of the data life cycle.**

Keywords—data management; stewardship; best practices; data sharing; data reuse

I. INTRODUCTION

Research data are valuable products of the scientific enterprise that historically have not been well preserved or archived. In recognition of this problem, research sponsors and scientific journals are increasingly encouraging or requiring sound data management, data preservation, and data sharing. Government agencies, for example, are under increasing pressure to demonstrate the benefits of the research they sponsor, both in terms of scientific findings (published papers) as well as data products. For instance, a 2007 US Government and Accounting Office Report summarized the issues associated with the loss of individual investigators' data and how this data loss deprives science and society of many of the benefits of research [1].

In January 2011, the National Science Foundation (NSF) instituted the requirement that a data management plan (up to two pages in length) be included as a supplement to every proposal [2]. Some individual NSF Directorates, Divisions, and Programs provide more specific guidelines; however, NSF is generally relying on scientists from the various disciplines it supports to set expectations for data management through the peer-review process. Educating the community about best data management practices is key to promoting a new culture of data stewardship, collaboration and data sharing.

In the remainder of this paper, we introduce DataONE and its approach to developing educational resources that promote good data stewardship. Next, we describe the Best Practices database and highlight, as a data management primer, a subset of the best practices that have been described to aid scientists in relation to all elements of the data life cycle. We conclude with recommendations for further development of educational resources that will benefit ecologists and environmental scientists.

II. DATAONE AND COMMUNITY ENGAGEMENT AND EDUCATION

DataONE is a federated data network that is being built to improve access to data about life on Earth and the environment that sustains it, and to support science by: (1) engaging the relevant science, data, and policy communities; (2) facilitating easy, secure, and persistent storage of data; and (3) disseminating integrated and user-friendly tools for data discovery, analysis, visualization, and decision-making.

A. Activities Central to DataONE

DataONE is being designed and built to provide a foundation for innovative environmental research that addresses questions of relevance to science and society. Five activities are central to the DataONE mission:

- Discovery and access: Enabling discovery and access to multi-scale, multi-discipline, and multi-national data through a single location.
- Data integration and synthesis: Assisting with the development of transformational tools that shape our understanding of Earth processes from local to global scales.
- Education and training: Providing essential skills (e.g., data management training, best practices, tool discovery) to enhance scientific enquiry.
- Building community: Combining expertise and resources across diverse communities to collectively

educate, advocate, and support trustworthy stewardship of scientific data.

- Data Sharing: Providing incentives and infrastructure for sharing of data from federally funded researchers in academia.

B. *Implementing DataONE*

Implementing the DataONE infrastructure requires that DataONE bring existing communities together in new ways. This is achieved via Community Engagement Working Groups that engage participants in identifying, describing, and implementing the DataONE cyberinfrastructure, governance, and sustainability models. These working groups, which consist of a diverse group of graduate students, educators, government and industry representatives, and leading computer, information, and library scientists:

- Perform computer science, informatics, and social science research related to all stages of the data life cycle;
- Develop DataONE interfaces and prototypes;
- Adopt/adapt interoperability standards;
- Create value-added technologies (e.g. semantic mediation, scientific workflows, and visualization) that facilitate data integration, analysis, and understanding;
- Address socio-cultural barriers to sustainable data preservation and data sharing; and
- Promote the adoption of best practices for managing the full data life cycle.

Community engagement and education activities are central to the DataONE mission. Activities designed to engage the community include: active participation of a diverse array of experts in DataONE Cyberinfrastructure and Community Engagement Working Groups; involvement of stakeholders from the international community in the DataONE Users Group, which meets annually; and numerous communication mechanisms including newsletters, Twitter, Facebook, and list serves. Education activities include two-hour to day-long training programs (e.g. "data management planning", "managing data for your research project") that are held at professional society meetings, webinars, three-week long graduate training in environmental information management, and the creation of education resources that include a tools database that highlights software tools that support all aspects of the data life cycle and a similar best practices database that is discussed below. A complete overview of DataONE, including working group activities, is currently in press for the *Journal of Ecological Informatics*.

III. BEST PRACTICES DATABASE

The best practices database was developed by 40 individuals that participated in two workshops. The first workshop was held in Santa Fe, New Mexico June 28-30, 2010, and resulted in a database consisting of 33 best practices. A second workshop was held in Santa Fe, New Mexico May 10-12, 2011 and resulted in the addition of 53 best practices for a current total of 86 database entries. Best practices were recommended by workshop participants based on experiences within their organizations and were revised and agreed upon by the other workshop participants.

The best practices database [3] consists of two related components. First, database entries consist of individual best practices. Individual entries include: the title of the best practice; the category of best practices to which the entry belongs; a brief phrase or sentence that summarizes the best practice; a complete description of the best practice that frequently includes examples; a rationale that highlights the benefits derived from employing the best practice; and additional information such as references to articles, books, or web sites where an individual can discover more detailed information. Box 1 provides an example of one of the best practices, "Assign descriptive file names". The overall database was designed to be easily searchable and the best practices have been condensed to short one-page descriptions. This was done to make it easy for scientists and students to rapidly answer individual questions they may have about managing their data without having to search through a book or lengthy technical documents.

The second component is a data management primer that is published for the first time below. The primer describes fundamentals of data management for scientists and students and highlights a subset of the specific best practices that are included in the database. The primer enables first-time users to get a comprehensive overview of good community practices as well as an understanding of the types of best practices they can expect to discover in searching the database.

IV. DATA MANAGEMENT PRIMER

Although data management plans may differ in format and content, several basic elements are central to effectively managing data. Ideally, data should be managed so that any scientist (including the collector or data originator) can discover, use, and interpret the data after a period of time has passed. A key component of data management is the comprehensive description of the data and contextual information that future researchers need to understand and use the data. This description is particularly important because the natural tendency is for the information content of a data set or database to undergo entropy over time (i.e. data entropy), ultimately becoming meaningless to scientists and others [4].

An effective data management program would enable a user 20 years or longer in the future to discover, access, understand, and use particular data [5]. This primer summarizes the elements of a data management program that would satisfy this 20-year rule. Specifically, it includes guidance on how to properly manage data, as well as how to effectively create, organize, manage, describe, preserve and share data—activities that are necessary to prevent data entropy.

Here we present a series of best practices that will help scientists manage the data they collect. We provide a guide on data management practices that investigators could perform during the course of data collection, processing, and analysis (components of the data life cycle, Fig. 1) to improve the

BOX 1. EXAMPLE OF AN ENTRY IN THE BEST PRACTICES DATABASE.

Title: Assign descriptive file names

Category: Data Files and File Management

Summary: File names should be descriptive and reflect the file content.

Best Practice: File names should reflect the contents of the file and include enough information to uniquely identify the data file. File names may contain information such as project acronym, study title, location, investigator, year(s) of study, data type, version number, and file type. Descriptive file names should not be a substitute for a complete metadata record.

When choosing a file name, check for any database management limitations on file name length and use of special characters. Also, in general, lower-case names are less software and platform dependent.

If versioning is desired a date string within the file name is recommended to indicate the version.

Avoid using file names such as mydata.dat or 1998.dat.

An example of a good data file name: Sevilleta_LTER_NM_2001_NPP.csv

Sevilleta_LTER is the project name

NM is the state abbreviation

2001 is the calendar year

NPP represents Net Primary Productivity data

csv stands for the file type—ASCII comma separated variable

Description Rationale: Clear, descriptive, and unique file names may be important when your data file is combined in a directory or FTP site with your own data files or with the data files of other investigators. File names that reflect the contents of the file and uniquely identify the data file enable precise search and discovery of particular files.

Additional Information:

Hook, L.A., S Santhana Vannan, T.W. Beaty, R.B. Cook, and B.E.Wilson. 2010. Best Practices for Preparing Environmental Data Sets to Share and Archive. Available at daac.ornl.gov/PI/BestPractices-2010.pdf. Oak Ridge National Laboratory Distributed Active Archive Center

Borer et al. 2009. Some Simple Guidelines for Effective Data Management. Bulletin of the Ecological Society of America 90: 209-214.

usability of their data. We assembled the most important practices that researchers could implement to make their data sets ready to share and to be re-used. These practices could be performed at any time during the preparation of the data set, but we suggest that researchers consider them in the data management planning stage, before the first measurements are taken.

A. Planning for Data Management

Plan for data management as your research proposal is being developed, whether development is for a funding agency proposal, a dissertation proposal, or some other project. The following should be considered:

1) Creating your data: Based on the hypotheses and sampling plan, what data will be generated? How will the samples be collected and analyzed? Provide descriptive documentation of collection rationale and methods, analysis methods, quality assurance methods, and any relevant contextual information.

2) Organizing your data: Decide on how data will be organized within a file, what file formats will be used, and the overall contents of the data products you will generate.

3) Managing your data: Who is in charge of managing the data? How will version control be handled? How will data be backed up, and how often?

4) Describing your data: Information that describes data is called metadata. How will you produce a metadata record? Which metadata standard will be used? What tool will you use? Will you create a record at the project inception and update it as you progress with your research? Where will you deposit the metadata?

5) Sharing your data: Develop a plan for sharing data with the project team, with other collaborators, and with the broader science community. Under what conditions will data be released to each of these groups? What are the target dates for release to these groups? How will the data be released?

6) Preserving your data: As files are created, implement a short-term data preservation plan that ensures that data can be recovered in the event of file loss (e.g. backing up data by storing the files routinely in several locations). Identify an appropriate long-term archive or database early in your project, and research that archive's requirements for data, documentation, and metadata.

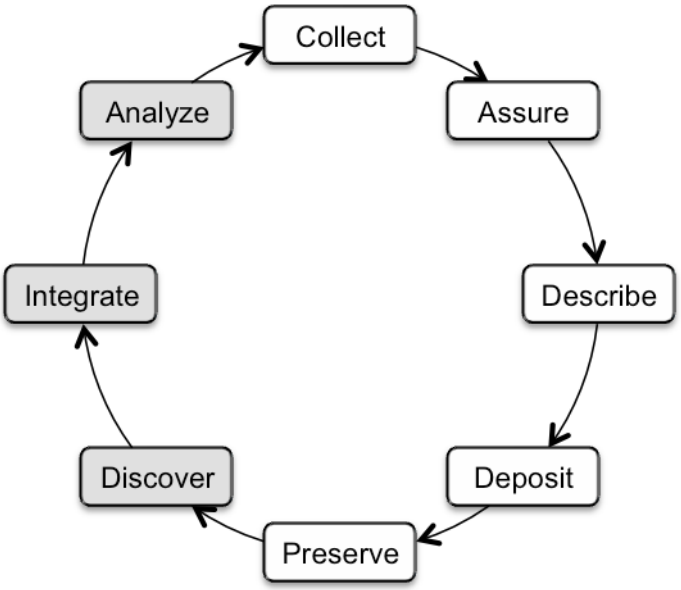


Figure 1. Data management life cycle from the perspective of a researcher. The entire life cycle comprises the key elements of a data management plan. The white boxes represent the steps an observational scientist takes to generate a primary data set for long-term archival, while the grey boxes represent the steps a data user may take to discover, integrate, and analyze existing data. A researcher can be both an observational scientist and a data user.

B. Managing Data Throughout the Data Life Cycle

A scientist or team of scientists is frequently engaged in all aspects of the data life cycle, both as a data creator and as a data user. Some scientists or teams (e.g. those engaged in modeling and synthesis) may create new data in the process of discovering, integrating, analyzing, and synthesizing existing data. This section summarizes best practices [6,7,8] for preparing data that can be readily shared with others.

1) Practices for Data Collection

a) Collect: Define the Contents of Your Data Files: Define each parameter, including its format, the units used, and codes for missing values. Provide examples of formats for common parameters.

b) Collect: Use Consistent Data Organization: We recommend that you organize the data within a file in one of two ways. Whichever style you use, be sure to place each observation on a separate line (row). In the first way to organize data, each row in a file represents a complete record and the columns represent all the parameters that make up the record (a spreadsheet format). In the second way, one column is used to define the parameter and another column is used for the value of the parameter (a database format). Other columns may be used for site, date, treatment, units of measure, etc. For specific examples, refer to [7].

c) Collect: Use Consistent File Structure and Stable Formats: Use the same format throughout the file – don't re-arrange columns or rows within the file. At the top of the file, include one or more header rows that identify the parameter and the units for each column. File formats should ideally be non-proprietary (e.g. .txt or .csv files rather than .xls), so that they are stable and can be read well into the future.

d) Collect: Assign Descriptive File Names: File names ideally describe the project, file contents, location, and date, and should be unique enough to stand alone as file descriptions. File names do not replace complete metadata records.

e) Assure: Perform quality assurance and quality control: check the format of the data to be sure it is consistent across the data set. Perform statistical and graphical summaries (e.g. max/min, average, range) to check for questionable or impossible values and to identify outliers. Communicate the quality of the data using either coding within the data set that indicate quality, or in the metadata.

f) Describe: Assign Descriptive Data Set Titles: When giving titles to data sets and associated documentation, be as descriptive as possible, because these data sets may be combined with other data sets and accessed many years in the future by people who will be unaware of the details of the project. Data set titles should contain the type of data and other information such as the date range, the location and, if applicable, the instruments used.

g) Describe: Provide Documentation: Comprehensive documentation is the key to future understanding of data. Without a thorough description of the context in which the data were collected, the measurements that were made, and the quality of the data, it is unlikely that the data can be easily discovered, understood, or effectively used. Use a stable file format to write your documentation (e.g. .html, .pdf, .txt) and refer to a specific data file. Both data and documentation

should have similar names (file names and titles). The documentation should describe what future researchers need to know to understand and use the data: the what, how, when, where, who, and additional contextual information for the study and observations.

h) Describe: Generate Metadata: Metadata should be generated in a format commonly used by the most relevant science community. Use metadata-editing tools (e.g. Metavist [9], Mercury Metadata Editor [10], Morpho [11]) to generate comprehensive descriptions of the data. Comprehensive metadata enable others to discover understand and use your data. Metadata should describe provenance of the data (where it originated, as well as any transformations the data underwent) and how to give credit for (cite) the data products.

i) Deposit: Work with a data center or archiving service that is familiar with the appropriate scientific domain. They will have a basic understanding of the data and can provide guidance as to how to prepare formal metadata and data set documentation, how to preserve the data, and how to provide additional services to future users of your data (discovery, access, integration, visualization, and analysis).

j) Preserve: During data collection, data should be secured and maintained, including performing regular backups. Ultimately, data should be preserved in a Data Center or archive that supports polices, procedures, and systems that protect the data. For appropriate attribution and provenance of a dataset, the following information should be included in the data documentation or the companion metadata file:

- The personnel responsible for the dataset throughout the lifetime of the dataset
- The context of the dataset with respect to a larger project or study (including links and related documentation), if applicable
- Revision history, including additions of new data and error corrections
- Links to source data, if the data were derived from another dataset
- Project support (e.g. funding agencies, collaborators, material support)
- How to properly cite the dataset

2) Practices for Ensuring Data Discovery and Reuse

a) Discover: Based on information submitted with the data (metadata), data centers can provide tools that support data discovery, access, and dissemination of data in response to users' needs. Use standard terminology and keywords to ensure that data can be searched for and discovered.

b) Integrate and Analyze: A variety of tools are available that support data integration, analysis, and visualization. Significant recent advances have been made in supporting the creation and management of complex, scientific workflows that serve to integrate, analyze, and visualize data as well as document the exact steps used in those processes [e.g. 12, 13, 14]. When datasets and data elements are used as a source for new datasets, it is important to identify and document those data within the documentation of the new derived dataset (i.e. provenance). This will enable (1) tracing the use of datasets and data elements, (2) attribution to the creators of the original datasets, and (3) identifying impacts of errors in the original datasets or data elements on derived datasets.

V. CONCLUSION

Data represent important products of the scientific enterprise that are, in many cases, of equivalent or greater value than the publications that are originally derived from the research process. For example, addressing many of the grand challenge scientific questions increasingly requires collaborative research and the re-use, integration, and synthesis of data. Consequently, academic, research and funding institutions are now requiring that scientists provide good stewardship of the data they collect. By implementing good data management practices early in the data life cycle, scientists can ensure that they are well prepared to meet these requirements.

The DataONE Best Practices Database represents an initial effort to educate scientists about best practices they can follow in managing their data. The database and accompanying primer (this paper) will continue to be updated in response to community feedback, as well as the availability of new enabling technologies. Further creation and refinement of educational resources such as the database and primer are important for enabling good data stewardship. However, these represent just one facet of the comprehensive education effort that is needed. In particular, we encourage professional societies to include data and information management training as a routine part of societal meetings because of the constant change in technology and the evolving expectations of research sponsors and the public. More importantly, we recommend that data management best practices be incorporated in introductory biology, ecology, and environmental science courses as well as in stand-alone graduate courses on data management. Such sociocultural changes are necessary if the next generation of scientists is to be equally knowledgeable of current scientific information as well as the data and informatics practices that lead to information and knowledge.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation (grant numbers 0753138 and 0830944) and National Aeronautic and Space Administration Grant NNG09HP121. We thank the many individuals who participated in the two workshops; their names and institutions are listed at www.dataone.org/dataonepedia.

REFERENCES

[1] United States Government Accountability Office Report, Accessed 5 June 2011. www.gao.gov/products/GAO-07-1172, 2007.

[2] "NSF Summary of Dissemination and Sharing of Research Results," Accessed 5 Jun 2011. http://www.nsf.gov/bfa/dias/policy/dmp.jsp

[3] "DataONEpedia: a DataONE online resource for Best Practices and Tools for Data Management", Accessed 5 June 2011. https://www.dataone.org/dataonepedia

[4] W. Michener, J. Brunt, J. Helly, T. Kirchner, and S. Stafford, "Nongeospatial metadata for the ecological sciences," Ecol. Applic., 7, pp. 330–342, 1997.

[5] National Research Council, "Solving the Global Change Puzzle: A U.S. Strategy for Managing Data and Information", Report by the Committee on Geophysical Data of the NRC Commission on Geosciences, Environment and Resources. National Academy Press, Washington, D.C., 1991.

[6] R. Cook, R. Olson, P. Kanciruk and L. Hook, "Best Practices for Preparing Ecological Data Sets to Share and Archive," Bull. Ecol. Soc. of Amer., vol. 82, 2001.

[7] L. Hook, S. Santhana Vannan, T. Beaty, R. Cook, and B.Wilson, "Best Practices for Preparing Environmental Data Sets to Share and Archive," Oak Ridge National Laboratory Distributed Active Archive Center. Accessed 5 June 2011. daac.ornl.gov/PI/BestPractices-2010.pdf, 2010.

[8] E. Borer, E. Seabloom, M. Jones, and M. Schildhauer, "Some Simple Guidelines for Effective Data Management," Bull. Ecol. Soc. Amer., vol. 90, pp. 209-214, 2009.

[9] Metavist Metadata Editor, Accessed 5 June 2011. http://metavist2.codeplex.com/

[10] Mercury Metadata Management, Data Discovery, and Access System, Accessed 5 June 2011. http://mercury.ornl.gov/

[11] Morpho Data Management Software Portal, Accessed 5 June 2011. http://knb.ecoinformatics.org/morphoportal.jsp

[12] B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger-Frank, M. Jones, E. Lee, J. Tao, and Y. Zhao, "Scientific workflow management and the Kepler system," Concurrency and Computation: Practice & Experience, vol. 18, 2006.

[13] T. Oinn, M. Greenwood, M. Addis, J. Ferris, K. Glover, C. Goble, D. Hull, D. Marvin, P. Li, and P. Lord, "Taverna: Lessons in creating a workflow environment for the life sciences," Concurrency and Computation: Practice and Experience, vol. 18, pp. 1067–1100, 2006.

[14] C. Goble, and D. DeRoure, "The impact of workflow tools on data-centric research," from The 4th Paradigm Part 3: Scientific Infrascturcture. Microsoft Research, Redmond WA, 2009, pp. 137-145.

Data Interoperability in the Hydrologic Sciences

The CUAHSI Hydrologic Information System

David G Tarboton[1], David Maidment[2], Ilya Zaslavsky[3], Dan Ames[4], Jon Goodall[5], Richard P Hooper[6], Jeff Horsburgh[1], David Valentine[3], Tim Whiteaker[2], Kim Schreuders[1]

[1] Utah State University
[2] University of Texas at Austin
[3] San Diego Supercomputer Center
[4] Idaho State University
[5] University of South Carolina
[6] Consortium of Universities for the Advancement of Hydrologic Science, Inc.

dtarb@usu.edu, maidment@mail.utexas.edu, zaslavsk@sdsc.edu, dan.ames@isu.edu, goodall@cec.sc.edu, RHooper@cuahsi.org, jeff.horsburgh@usu.edu, valentin@sdsc.edu, twhit@mail.utexas.edu, kim.schreuders@usu.edu

***Abstract*—The Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) has been established to promote research infrastructure that advances Hydrologic Sciences. Hydrologic Information Systems (HIS) are part of this infrastructure. Hydrologic information is collected by many individuals and organizations in government and academia for many purposes, including general monitoring of the condition of the water environment and specific investigations of hydrologic processes and environments. This paper describes HIS capability developed to promote data sharing and interoperability in the Hydrologic Sciences with the ultimate goal of enabling hydrologic analyses that integrate data from multiple sources. The CUAHSI HIS is an internet based system to support the sharing of hydrologic data. It is comprised of hydrologic databases and servers connected through web services as well as software for data publication, discovery and access. The system that has been developed provides new opportunities for the water research community to approach the management, publication, and analysis of their data systematically. The system's flexibility in storing and enabling public access to similarly formatted data and metadata has created a community data resource from public and academic data that might otherwise have been confined to the private files of agencies or individual investigators. HIS provides an analysis environment for the integration of data from multiple sources and serves as a prototype for the infrastructure to support a network of large scale environmental observatories or research watersheds.**

Keywords—Hydrologic Information System; Web services; Data Model; Hydrology

I. INTRODUCTION

The advancement of hydrologic science is critically dependent on the assembly and synthesis of hydrologic data. The Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI) is an organization representing 135 universities and affiliated organizations, funded by the US National Science Foundation, to develop community infrastructure and services to advance hydrologic science. This paper describes the CUAHSI Hydrologic Information System (HIS), a community information systems technology project to improve access to hydrologic data.

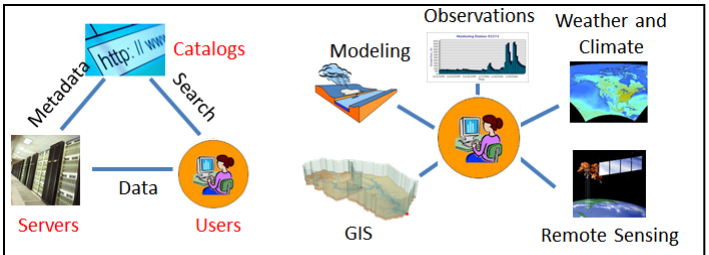
The CUAHSI HIS project [1, 2] has as a goal the development of standards, systems, and software to enhance access to and interoperability among water data from multiple sources. We have built a prototype system centered on a services-oriented architecture [3] that defines the interfaces between system components for publishing, cataloging and accessing hydrologic data and a desktop hydrologic information system that supports the integration and analysis of hydrologic data retrieved from multiple sources.

II. ARCHITECTURE

Two concepts, (1) the services oriented architecture; and (2) the desktop hydrologic information system underlie the architecture of the system that we are developing (Fig. 1).

The HIS services-oriented architecture can be viewed as: 1) a way of publishing hydrologic data in a uniform way; 2) a way of discovering and accessing remote water information archives in a uniform way; and 3) a way of displaying, synthesizing and analyzing water information and exporting it to other analysis and modeling systems. The connections among components are established by web services.

The concept of HIS desktop application software is somewhat analogous to Geographic Information System (GIS) desktop software that supports storage and analysis of logically linked data [4]. Our implementation, "HydroDesktop" provides an analysis environment within which data from multiple sources can be discovered, accessed and integrated.



a) Services-Oriented Architecture b) Desktop Hydrologic Information System

Figure 1. Hydrologic Information System Overarching Vision.

We have developed prototype functionality for all three components of the services oriented architecture and data transmission formats for the data exchanges between them. In terms of the desktop hydrologic information system, we have developed a prototype desktop application that combines the analysis of GIS, modeling and observations. It downloads, stores and operates on the information on a local desktop computer. Our present implementation is still under active development and has not yet developed the capability to integrate weather, climate and remote sensing data illustrated in Fig. 1, but does synthesize GIS, point observations and time series and modeling.

The HIS services-oriented architecture is comprised of three classes of functionality: 1) data publication (HydroServer), 2) data cataloging (HydroCatalog), and 3) data discovery, access and analysis (HydroDesktop) (Fig. 2). This functionality follows the general paradigm of the Internet. HydroServer publishes data similar to the way Internet web servers publish content. HydroDesktop consumes data published from HydroServer, similar to the way web browsers consume Internet content. HydroCatalog supports data discovery based on indexed metadata similar to the way search engines support the discovery of Internet content.

The components shown in Fig. 2 either publish or consume information via the following categories of web services:

- Data Services – which convey the actual data.
- Metadata Services – which convey metadata about specific collections or series of data.
- Search Services – which enable search, discovery, and selection of data and convey metadata required for accessing data using data services.

The formats for transmission of information between these systems and the interfaces that enable the communication between them (the connecting arrows in Fig. 2) are critical to the functioning of the system. CUAHSI HIS has developed WaterML, an XML based language for transmitting observation data via web services [5]. The web services are referred to as WaterOneFlow web services. CUAHSI HIS also relies on other established standards such as World Wide Web Consortium Simple Object Access Protocol (SOAP) and Open Geospatial Consortium (OGC) Geographic Markup Language (GML) for transmission of information between the three primary components.

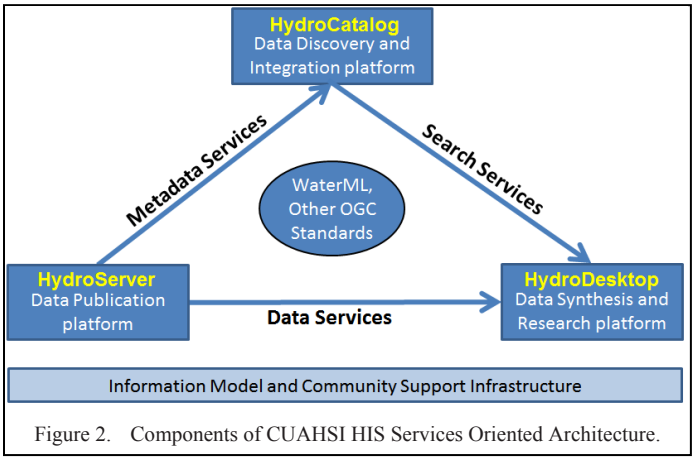


Figure 2. Components of CUAHSI HIS Services Oriented Architecture.

At the base of Fig. 2 is the information model and community support infrastructure upon which the system is founded. The information model is the conceptual model used to organize and define sufficient metadata about hydrologic observations for them to be unambiguously interpreted and used. Within HydroServer, it is encoded using the Observations Data Model (ODM) [6] relational database and the HydroServer Capabilities Database to ensure that data and metadata are stored together. The information model also serves as the conceptual basis for WaterML to ensure that data and associated metadata are transmitted with fidelity when data are downloaded. HydroDesktop implements the information model within its data repository database ensuring that local copies of data retrieved from a server maintain their original context. ODM includes a number of controlled vocabularies for metadata such as units, variable names, sample media etc., where semantic consistency in describing observations is important. The information model also includes a defined ontology used to represent a hierarchy of concepts that categorize the variables being measured. The ontology has been developed to support concept based search. The ontology and controlled vocabulary components of the information model have been developed to provide semantic consistency of the terms used in metadata and to support search and discovery based on these semantics. A web site collects and manages community additions and edits to controlled vocabulary content to allow dynamic growth of this content while encouraging semantic consistency across the user community.

The architecture shown in Fig. 2 has evolved as an approach for sharing hydrologic observations data that is general and open to allow broad participation. The HydroServer software stack is not the only entry point for data publishers. Anyone can publish data using web services that deliver data in WaterML format and thus have their data become part of this system. Similarly the HydroCatalog and HydroDesktop functionality is not limited to the software we have developed. The definition of standard functionality for transmission of information to and from a catalog provider enables others to establish their own catalogs. HydroDesktop is our prototype client for consumption of web service based hydrologic data, but this does not preclude others from establishing their own clients.

III. HydroServer

HydroServer is envisioned to be a self-contained, complete hydrologic data and metadata publication system that permits data publishers to control their own data while still being part of a distributed national/international system allowing universal access to the data [7]. HydroServer is targeted at investigators who are collecting data within research watersheds or observatories, although the software is general and can be used by anyone who wants to share hydrologic observations. The HydroServer software stack relies on the protocols and standards established by the HIS project and consists of a number of software applications that are being developed and managed as open source software using an open source code repository (http://hydroserver.codeplex.com).

An important principle that has emerged from our work on HydroServer , is that server functionality should support complete description of the data and metadata. We refer to this as the self-describing principle and this stems from the fact that the person or organization creating the data is generally best suited to provide metadata, and should have control over data publication. A catalog should not be required to aquire or generate additional metadata when supporting the discovery of data from a HydroServer.

HydroServer (Fig. 3) supports publication of both point observations data stored in one or more ODM databases [6] and published using WaterOneFlow web services and geospatial data published using OGC Web services from ArcGIS Server. Each HydroServer has a Capabilities Database that catalogs metadata about the data and web services it publishes. The Capabilities Web Service includes methods that return, in XML format, the list of regions for which data have been published, the published point observations data services, and the list of published spatial data services, along with appropriate metadata for each. By doing so, all of the capabilities of the HydroServer can be discovered and metadata harvested automatically by registration and cataloging services (HydroCatalog), making a HydroServer self-describing. These three web services comprise the service interface.

A suite of tools to load, edit and assist with the management of ODM data has been developed. A configuration tool has been built that provides an interface for defining the contents of the Capabilities Database. The ODM Tools suite and capabilities configuration tool comprise the data manager interface.

Finally, a suite of data presentation and visualization tools has been created for HydroServer. The suite includes the HydroServer Website, the Time Series Analyst, and the HydroServer Map Website. These provide a public browser accessible graphical user interface to the data holdings of the HydroServer.

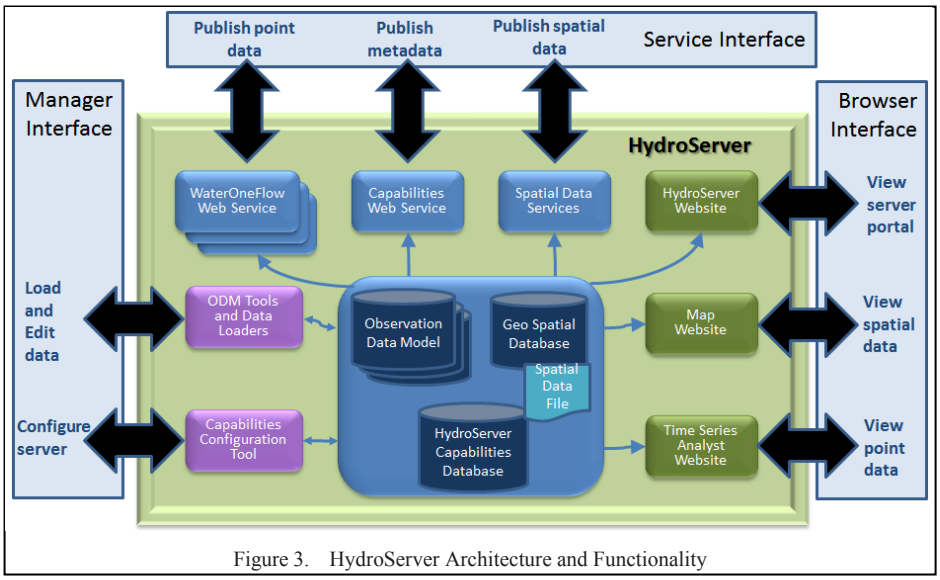


Figure 3. HydroServer Architecture and Functionality

IV. HydroCatalog

HydroCatalog is the discovery component of the system linking data publishers and application clients. Data discovery across multiple data services is enabled by a centralized Metadata Catalog Database, which contains descriptions of the datasets hosted on the many federated data servers on which data are published. HydroCatalog interfaces with data publishers through its web sites, interfaces with WaterOneFlow web services, and interfaces with desktop clients through search and ontology web services (Fig. 4).

HydroCatalog supports discovery of data by keywords, which represent concepts in the CUAHSI ontology and a collection of their synonyms. Search functionality requires that variable names in registered services are associated with terms at the nodes of this hierarchy. Data publishers first register their WaterOneFlow web services with the HydroCatalog Web Service Registry. Registration of a service triggers the Metadata Harvester to harvest the metadata from the web service and store it in the metadata catalog database. Once the metadata is stored in the database, data publishers can use the tagging application on the Semantic Annotation Website to map their variables to terms in the hydrologic ontology. The ontology can be visualized on part of the Semantic Annotation website (currently at http://hiscentral.cuahsi.org/startree.aspx).

Once tagging is complete, the metadata are discoverable through the Search and Ontology Web Service. The metadata harvester does periodic metadata harvests for each of the registered WaterOneFlow web services to ensure that the metadata catalog database is kept up to date. A Logging Service records use information on WaterOneFlow services that report use back to HydroCatalog. The Monitoring Service periodically accesses registered WaterOneFlow services to monitor their status so that breaks in service may be identified

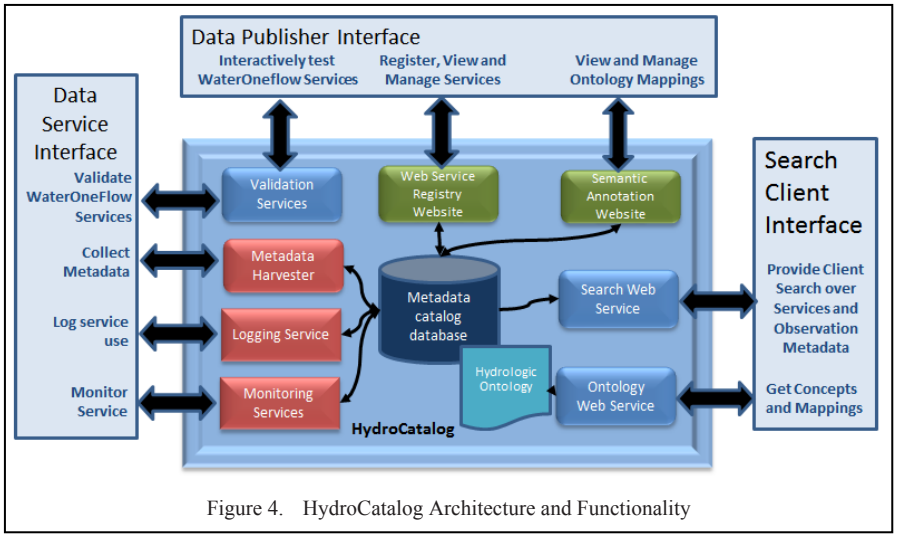


Figure 4. HydroCatalog Architecture and Functionality

and rectified, or services that go offline be de-listed (after first attempting to work with their owners to reinstate them).

The Search and Ontology Web Service that exposes the contents of the metadata catalog database includes a number of web service methods that enable spatial, temporal, and semantic searches across all sources of data in the catalog. Search results contain all of the information necessary to retrieve data in WaterML format from the data server on which the data are hosted, and client applications that use the HydroCatalog search services (e.g., HydroDesktop) can use the information contained within the search results to retrieve the data on demand. HydroCatalog software is open source software managed at (http://hydrocatalog.codeplex.com).

V. HydroDesktop

HydroDesktop is a free and open source Desktop Hydrologic Information System (Fig. 5) that helps users discover, use, manage, analyze and model hydrologic data.

The Geographic Information System (GIS) components of HydroDesktop are built from the open source DotSpatial library, while the time series components use HIS web services. The result is a spatially-enabled system for downloading observational data describing our water environment. The architecture of HydroDesktop (Fig. 5) is structured to take advantage of centralized cataloging functionality from HydroCatalog as well as distributed data from HydroServers.

The DotSpatial project (http://dotspatial.codeplex.com/) has been under development by members of the HIS team as well as an international open source volunteer community and members of the MapWindow project (see mapwindow.org) since April 2010. Since it's first release, DotSpatial has been downloaded over 40,000 times and it currently receives approximately 200 downloads per day by user-developers exploring free and open source alternatives for GIS enabled custom software targeting the Microsoft Windows operating system.

The DotSpatial engine used by HydroDesktop provides geographic visualization capability. HydroDesktop uses a plugin architecture, and plugins support searching for, downloading, viewing, graphing, editing, exporting, printing, and modeling with time series data. The search plugin allows search by area, time range, key words, and server. Like HydroServer, HydroDesktop is open source software developed using an open source code repository (http://hydrodesktop.codeplex.com).

At the heart of HydroDesktop is the capability to search for, discover, download, visualize and export data from the HIS network. Search and discovery is primarily achieved through a search plugin that allows a user to search based on:

- Area – The user must select a polygon on the map from one of the default data layers (counties, states, major watersheds) or from a polygon layer added by the user. Alternatively the user can draw a box on the map to identify a search area.
- Key Words – The user can optionally specify a set of key words related to observed variables to be used in the search query. Key words can be found by browsing a tree-view control or by typing key words in a search box. If no key words are selected then the query defaults to all variables.
- HydroServers – The user can optionally specify specific HydroServers or HIS services to include in the query. If none are specified then all known services are included in the search.
- Time Range – The user can optionally specify a time range for the data search by indicating a start and stop date which bound the time period of interest.

The user creates the search and executes it. This results in the creation of a "search results" layer showing all points on the map where data series were found. The user then selects series of interest from the map and executes a data download function which retrieves all of the data to the local computer database.

Once data have been downloaded into the HydroDesktop database, they can be immediately viewed graphically or

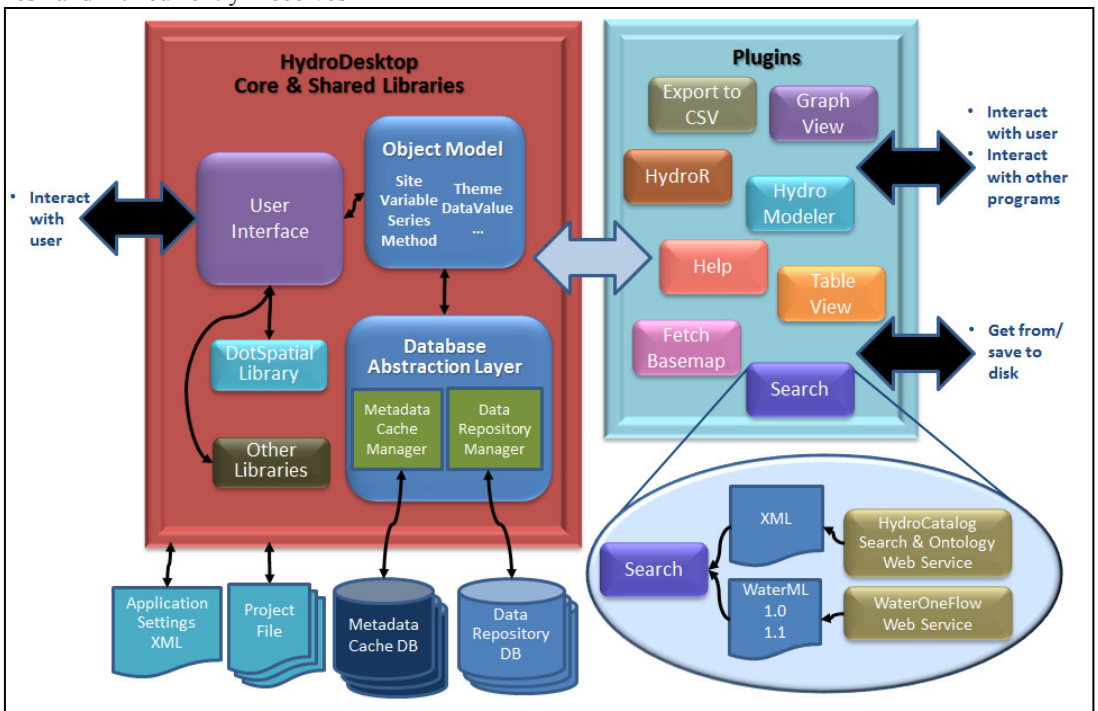


Figure 5. HydroDesktop Architecture and Functionality

tabularly through a "Graph View" plugin and a "Table View" plugin respectively. Graph visualization includes the ability to view time series, probability, histogram, and box-and-whisker plots that are extensively customizable and can be exported as graphic files for use in reports or other documents. The Table View plugin allows the user to view the data in tabular form and export the data to a comma separated values (CSV) file. The "Edit View" plugin enables editing.

Through its plugin interface, HydroDesktop has been extended to support extensive statistical analysis and modeling capabilities. Recognizing the cost prohibitive challenges and associated massive software development effort that would be required to build custom statistical analysis and modeling capabilities natively into the HydroDesktop application, HIS project team members made the decision early in the project to provide such capabilities through coupling with 3rd party software applications. Specifically two unique and very powerful plugins have been constructed for HydroDesktop. The first is a plugin called HydroModeler that leverages the OpenMI modeling framework developed under European Union funding. OpenMI (see www.openmi.org) defines a model interoperability interface that allows hydrologic and other time-step based models to interact with each other – passing data between models – as needed to simulate complex natural systems. The HydroModeler plugin to HydroDesktop provides an implementation of the 1.4 version of the OpenMI standard and specifically allows modelers to read HIS derived datasets into their models and write model outputs back into the HydroDesktop database.

The second 3rd party software which has been wrapped in the HydroDesktop plugin environment is the statistical software, "R". R is an extremely powerful script/command line based open source statistical analysis software tool based on the same scripting language used in the popular proprietary "S-Plus" software. The HydroR plugin provides an R scripting and execution environment directly within HydroDesktop, thereby extending the statistical analysis capabilities of HydroDesktop immensely. Fig. 6 illustrates the HydroDesktop interface highlighting the integration of data from multiple sources and combining, map, graph and search capability.

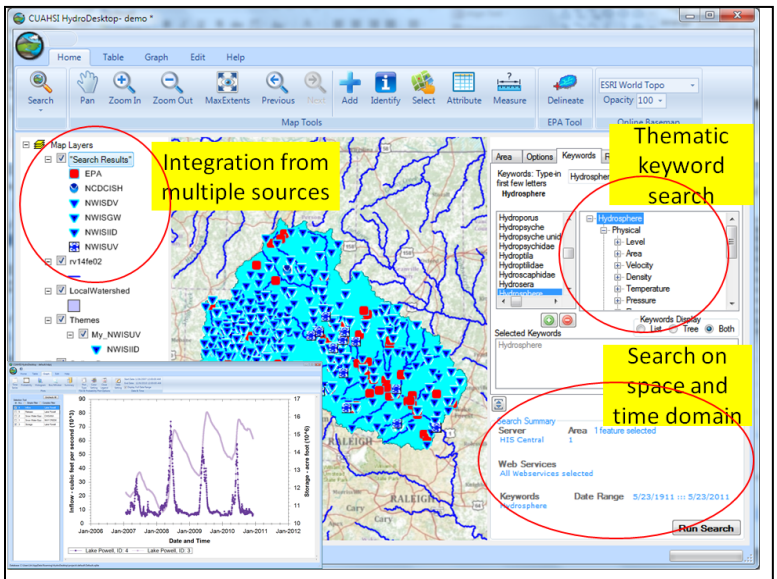


Figure 6. HydroDesktop Interface Illustration

VI. Use and Community Support

Table 1 summarizes the data available and its recorded use from instances of HydroServer registered with the CUAHSI HydroCatalog at SDSC. There is also use of the open source software that is downloaded by others and not registered here for which we do not have data. Standard HydroServer refers to installations, typically at universities, that have used the HydroServer software stack we have developed to publish data. Hybrid HydroServer refers to large existing federal datasets that the HIS project has wrapped with a WaterOneFlow web service.

The United States Geological Survey (USGS) and the National Climatic Data Center (NCDC) have adopted WaterML for publication of some of their data and have programmed web services that support some of the HydroServer functionality from their systems. The USGS daily and instantaneous value services (http://waterservices.usgs.gov/rest/USGS-DV-Service.html and http://waterservices.usgs.gov/rest/WOF-IV-Service.html) provide data encoded as WaterML. Similarly, NCDC serves data in WaterML format for some of their climate data online datasets (http://www7.ncdc.noaa.gov/rest/). It is through broad uptake of the services oriented architecture of the HIS, based on existing and emerging standards, that this system will become sustainable.

TABLE I. CUAHSI HydroServer Use Data

	Standard HydroServer	Hybrid HydroServer
Number of registered WaterOneFlow data services	66	6
Number of sites	462,992	1,490,113
Number of variables	5,978	6,892
Number of data values	>4 billion	>0.9 billion
Number of GetValues requests 7/1/2009-6/30/2010	46,055	64,810[b]
Number of GetValues requests 7/1/2010-6/30/2011	571,560[a]	43,723[b]

a. 435,762 of these are from the new West Gulf River Forecast Center NEXRAD precipitation data service that started in the latter year.

b. These are dominated by USGS NWIS Unit Values requests that dropped off when services to obtain this data directly from the USGS became available.

Reliance on independently developed and governed standards is one of the key elements of project sustainability. Other considerations that support sustainability are:

- Interacting with the community of CUAHSI HIS adopters and users
- Cultivating an open software development model (including infrastructure to support distributed code management, code reviews and refactoring, unit and user interface testing, automated builds) and

encouraging contributions from developers outside the project team

- Education and dissemination effort (seminars, workshops, presentations, class exercises, tutorials, learning modules)
- Maintaining a solid operational foundation of the system (high availability data discovery system, hardware and service monitoring and reporting, service testing and validation)
- Engagement with key, long-standing government, university and industry groups, capable of contributing to the system and data development and maintenance beyond the funding cycle (federal and state agencies, libraries, leading companies such as ESRI and Kisters)
- Extending CUAHSI HIS technology in several NSF-supported research and cyberinfrastructure projects

Development of HIS is done under the auspices of CUAHSI with 135 member organizations (mostly university), which sets policies such as software licensing, data publication and data use agreements. CUAHSI is advised by its Informatics Standing Committee that provides user and community input on priorities and needs necessary to support the academic research community.

VII. Conclusions

There is a fundamental need within the hydrologic and environmental engineering communities for new, scientific methods to organize and utilize observational data that overcome the syntactic and semantic heterogeneity in data from different experimental sites and sources and that allow data collectors to publish their observations so that they can easily be accessed and interpreted by others. The tools and partnerships that CUAHSI HIS has developed provide: (1) **Data Storage** in an Observations Data Model (ODM) and publication through HydroServer; (2) **Data Access** through internet-based WaterOneFlow web services using a consistent data language, called WaterML from HydroDesktop; (3) **Data Discovery** through a National Water Metadata Catalog and thematic keyword search system at HydroCatalog and (4) **Integrated Modeling and Analysis** within HydroDesktop. These functions support a high level of interoperability for hydrologic data. Beyond technical aspects, HIS has also focused on scientific, organizational, and infrastructure aspects of hydrologic data integration, which represent an important part of its contribution – in particular building partnerships with major federal and state agencies to incorporate their data into the system and ingrate with data provided by multiple academic partners.

The HIS is a federated system linking data from multiple providers. As such, data availability and quality does depend on it being maintained by the provider. The ODM data model provides capability to document data sources, methods and quality controls, but there is no filter on data quality that may be published using HIS technology. In this respect the system is also like much other information on the internet, buyer beware, user's need to assess for themselves the suitability of data for a particular purpose. As with broken links on the internet, when servers go down data becomes unavailable. The system does enable the capability for institutions to establish data centers to store data that is critical to them and CUAHSI is working to establish such a long term data center to archive community data.

The combination of HIS capabilities creates a common window on water observations data for the United States unlike any that has existed before, and is also extensible worldwide. This system represents new opportunities for the water research community to approach the management, publication, and analysis of their data systematically. The system's flexibility in storing and enabling public access to similarly formatted data and metadata has created a community data resource from public and academic data that might otherwise have been confined to the private files of agencies or individual investigators. HydroDesktop provides an analysis environment for the integration of data from multiple sources and serves as a prototype for the infrastructure to support a network of large scale environmental observatories or research watersheds.

For more information about the CUAHSI HIS and access to the tools and code, all freely distributed and open source, under the Berkeley Software Distribution (BSD) license, go to our website: http://his.cuahsi.org.

Acknowledgment

Funding for this work, by the U.S. National Science Foundation under grant EAR 0622374 is greatly appreciated. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] D. G. Tarboton, D. Maidment, I. Zaslavsky, D. P. Ames, J. Goodall, and J. S. Horsburgh, "CUAHSI hydrologic information system 2010 status report," 2010. http://his.cuahsi.org/documents/CUAHSIHIS2010StatusReport.pdf.

[2] D. G. Tarboton, J. S. Horsburgh, D. R. Maidment, T. Whiteaker, I. Zaslavsky, M. Piasecki, J. Goodall, D. Valentine, and T. Whitenack, "Development of a community hydrologic information system," in *18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation*, 2009, pp. 988-994 http://www.mssanz.org.au/modsim09/C4/tarboton_C4.pdf.

[3] N. M. Josuttis, *SOA in practice - the art of distributed system design*. Sebastapol, CA: O'Reilly Press, 2007.

[4] R. Tomlinson, *Thinking about GIS*. Redlands CA: ESRI Press, 2003.

[5] I. Zaslavsky, D. Valentine, and T. Whiteaker, "CUAHSI WaterML," Open Geospatial Consortium Discussion Paper OGC 07-041r1, 2007. http://portal.opengeospatial.org/files/?artifact_id=21743.

[6] J. S. Horsburgh, D. G. Tarboton, D. R. Maidment, and I. Zaslavsky, "A relational model for environmental and water resources data," *Water Resour. Res.,* vol. 44, p. W05406, 2008. doi:10.1029/2007WR006392.

[7] J. S. Horsburgh, D. G. Tarboton, K. A. T. Schreuders, D. R. Maidment, I. Zaslavsky, and D. Valentine, "Hydroserver: A platform for publishing space-time hydrologic datasets," in *2010 AWRA Spring Specialty Conference Geographic Information Systems (GIS) and Water Resources VI*, Orlando Florida, 2010 http://www.awra.org/orlando2010/doc/abs/JefferyHorsburgh_7cb420e3_6602.pdf.

Understanding the data management needs and data sharing challenges of environmental scientists

Carol Tenopir[1], Suzie Allard[1], Miriam Davis[1] *(Authors)*
[1] The University of Tennessee
ctenopir@utk.edu, sallard@utk.edu, miriams@utk.edu

***Abstract*— Surveys of scientists for the NSF DataONE project and the USGS Southeast Information Node of the National Biological Information Infrastructure (NBII), as well as follow-up interviews, show that environmental scientists are interested in sharing their data with certain conditions, such as citations or acknowledgment. Government scientists are more likely to be satisfied with the processes for data management than are academic scientists, but less likely to be satisfied with the process of describing data or tools for documentation. Both groups value trusted and complete sources. There are many ways that scientists can be assisted with data management throughout the data life cycle.**

Keywords—data management, environmental information, information needs of scientists, data practices of scientists

I. Introduction

Access to data and information resources are critical to the work of science, yet environmental scientists cannot always access what they need and do not always know how to prepare their own data for long term sharing with others. Understanding current data management practices, as well as the needs, barriers and challenges of data management for the future, will help information system designers, librarians, informationalists, and data managers provide better services to scientists now and into the future [1] [2] [3] [4] [5] [6].

Baseline assessments are important because they provide understanding of the practices of a group at a fixed point in time. On-going assessments can be used to judge changes over time, providing a means of demonstrating improvement. An important first step in the assessment process is to better understand the needs and practices of scientists today.

Much of the work investigating scientists' information needs focuses on the research needed on particular topics, or in particular fields, and the analyses needed to address particular questions and issues within those topics and fields, rather than on information needs per se. Research needs and information needs are similar, but scientific research is just one input of information needs. Other information needs would include the types of information and information tools needed, the attributes of the most useful information, etc.

Furthermore, the information needs of user types differ. While research scientists tend to focus on research needs and characteristics or attributes of their topic of focus, such as ecosystems, environmental decision makers, including natural resource managers, are more likely to require integrated information and tools that highlight patterns and relationships between various factors, decision support tools, and the integration of scientific and social data [8]. In this sense, models are an important information source for environmental scientists. Appropriately scaled information has also been identified by many studies as a need of environmental scientists [8].

In terms of scientists' information practices, more studies are needed across the data life cycle from data acquisition through data management to data sharing, archiving and re-use. To date, data sharing and data management have garnered the most attention. Results of past studies indicate that while interest and support for data sharing, especially related to data generated by publicly funded research projects, is high, actual data sharing among scientists is minimal (although practices do vary across fields) [1], [2], [4]. Fields with cultures supportive of data sharing practices and attitudes within various subject disciplines have been studied, analyses of these factors among environmental scientists working in different scientific sectors such as academia or government appears to be missing. Regardless of discipline, among the many reasons for withholding data are amount of effort vs. payoff in terms of career interests or furthering knowledge in a particular field, preservation of ability to publish, and misuse of data [5], [9], [2]. Institutional policies and procedures can also be as great a barrier to data sharing, or greater, than individual scientific preference [1], [5], [2], [6].

For this project, surveys and interview assessments were conducted among biological and environmental scientists in 2010 and 2011 to help understand practices, needs, and challenges relating to research data management. We frame our assessments in terms of the complete Data Life Cycle—that is, all of the processes from data collection, quality assurance, metadata description, deposition into a trusted node, preservation, and then discovery, integration with other datasets, analysis, and once more collection of new data (see Fig. 1). Assessment of user practices, perceptions, and needs are essential throughout this process, to help build better products and to move discovery forward.

These efforts were part of two projects: 1) NSF Data Observation Network for Earth (DataONE) and 2) USGS Increasing Biodiversity Information Sources (IBIS). DataONE

is a large international project, led by Principal Investigator William Michener at the University of New Mexico. (For more information see www.dataone.org.) The University of Tennessee team was responsible for the assessments of needs and current practices surrounding research data. IBIS is a project at the University of Tennessee in support of the Southeast Information Node of the USGS National Biological Information Infrastructure. (www.nbii.gov.)

DataONE is designed to be the foundation of new innovative environmental research by ensuring preservation and access to multi-scale, multi-discipline, and multi-national data. DataONE is unique in that it: (1) builds on existing data repositories including data centers; (2) creates a global, federated data network by focusing on interoperability and providing tools and services to enable new science and knowledge creation; and (3) supports evolving communities of practice enabled by the DataONE cyberinfrastructure and informed by best practices, exemplary data management plans, and tools that support all aspects of the data life cycle. (www.dataone.org.)

The Usability & Assessment and Sociocultural Working groups of DataONE are responsible for baseline and ongoing assessments of all stakeholders. The focus of DataONE assessments started with its primary group of stakeholders – scientists, who were the priority group to inform all activities across DataONE.

IBIS was a three year project for the Southeast Information Node (SEIN) of USGS. The project focused on understanding the information and data needs of southeastern U.S. Scientists and facilitating access to high quality information sources and data sets. The efforts were aligned with USGS science priorities: first climate change, followed by aquatics and renewable energy as they pertain to biodiversity. A survey and interviews of southeastern scientists have provided us with insights into data practices and needs.

Both projects use assessments of scientists to gain insights into how scientists collect, use, share, and curate data and what tools and other support they need to make those processes better. The differences between the two projects include scope—international for DataONE versus southeast U.S. for IBIS, funding agency (NSF vs. USGS), and specific subject focus (earth and environmental sciences for DataONE and biodiversity for USGS). The ultimate goals of both DataONE and IBIS, however, are to enhance the practice of science through enabling data and information discovery that allows scientists to quickly respond to emerging environmental issues. The assessments also highlight partnerships that have been developed between DataONE and IBIS.

Figure 1. Data Lifecycle

The findings from these surveys will help USGS and NSF understand the issues and needs of scientists that will improve data management and data sharing. Improved access to data will help forward earth and environmental science discovery and collaborative science now and into the future.

II. FINDINGS

The DataONE survey was distributed via ―champions‖—that is volunteers from various institutions emailed the survey to their faculty and colleagues. From an estimated 9,000 invitations, over 1300 responses (1329) were received, mostly from across North America (73%) or Europe (15%). Most of the respondents were from academic institutions (80%), with 13% from government agencies. Biological, environmental, and ecology scientists were the largest number of respondents (>50% combined), but respondents also came from the social sciences, physical sciences, and other disciplines (see Fig.2).

The IBIS survey was much more focused—respondents came from eight states in the southeastern United States, with just over half from academic institutions. Email invitations were sent to science faculty at many research universities in the southeast and to employees of state, local and non-profit environmental agencies, with 428 total respondents (See Fig. 3). A large number of the government respondents are from federal agencies (69% of the government respondents). A majority of respondents came from life sciences (52%) and agriculture and natural resources (24%).

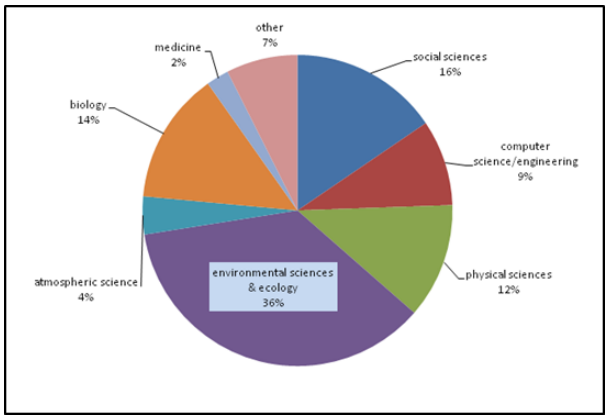


Figure 2. Subject disciplines (DataONE)

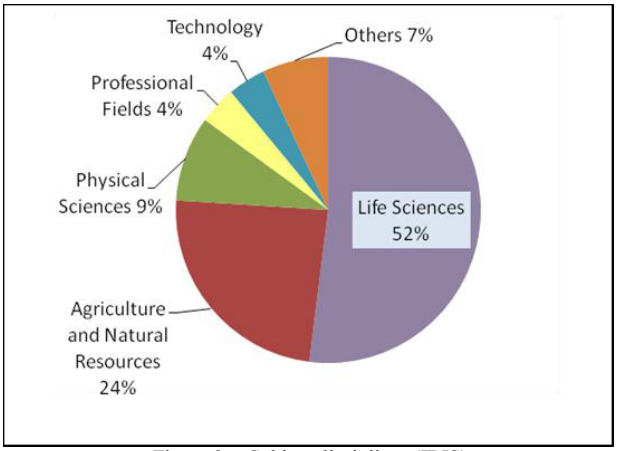


Figure 3. Subject disciplines (IBIS)

In-depth follow-up interviews with 30 southeastern scientists and data managers who are interested in sharing or preserving their datasets provided additional insights into the motivations and practices for data management of their own datasets. Approximately half of those interviewed worked in academic institutions, a quarter in government agencies, and a quarter in non-profit organizations. These interviews also identified the existence of unique data sets and are informing the development of personas, or characters created to represent typical users of science data and information products and services. Throughout this paper we refer to two of these personas—Joe, a biodiversity specialist employed in a government agency and Mabel, an academic environmental scientist. Joe and Mabel are typical representatives of the government and academic scientists we spoke with. Their quotes are drawn directly from the aggregated interviews.

The surveys and interviews asked many related, but different, questions. Many integrated lessons learned emerged by examining surveys and interviews together. Six of the lessons learned are explored in more detail in this paper:

1. Scientists need a variety of data types.
2. Many scientists are interested in sharing data.
3. There are many barriers to sharing data and conditions that must be met.
4. There are different needs, attitudes, and practices between scientists who work in government agencies and those who work in academia.
5. The skill level of scientists and use and access to appropriate tools varies across the data life cycle.
6. There are many ways that scientists can be assisted across the data life cycle.

Lesson 1: Scientists need a variety of data types

It may come as no surprise that the range of data types collected and used by scientists varies widely. Although experimental (54%) and observational (48%) data are the most frequently used, data models (38%), abiotic (34%) and biotic (33%) surveys from both field collection and remote sensors are also used. Since few of our respondents are social scientists it is not surprising that human subject surveys or interviews are less common, in particular among government scientists. This is not to say, however, that scientists, including government scientists, do not need access to social science data. In fact, a common theme from past studies, particularly among natural resource managers and other environmental decision makers, is the need for integrated science and social data for the purposes of decision making, as well as information and tools that summarize patterns and relationships. Our findings support these. In terms of the types of data needed to do their work, south eastern scientists need equal access to raw data (65%) and summarized data (65%). Over half (52%) say that data models are essential or important to their work.

These findings can provide guidance in prioritizing the development of information products and tools, so that efforts can be concentrated on those data tools that will serve the largest community of users.

Lesson 2: Many scientists are interested in sharing data

Gaining access to data is one part of the challenge, scientists being willing to share those data is another. At least three-quarters of all scientists surveyed say they currently share their data with others and 78% are willing to put at least some of their data into a central repository. Many fewer say they are willing to share ALL of their data, however. Only 41% are willing to share all of their data in a central repository with no restrictions.

The government scientist persona echoes those sentiments that sharing is ultimately for the good, with some concerns and need for some restrictions. Joe says:

–We are torn between putting it out there for everyone and worry about suffering the risk of something bad happening with it. Saddest thing would be if the data loses its use where it isn't shared.

–I don't think I would be opposed to it. It would not be a decision I would make personally; we would have to have permission to share.

Academic scientists are more obviously enthusiastic about data sharing and reuse. Mabel says:

–I'm interested in having data available to researchers interested in larger questions, particularly climate change questions.

-If NBII required anyone who extracted data through the portal to also share data with the portal, then a resounding yes.

Lesson 3: There are many barriers to sharing data and conditions that must be met.

Right now, only 36% agree that others can access their data easily, even though they may be willing to share some of their data by placing it into a central repository. This gap between willingness to share and perceived accessibility of their data reflects past findings from the literature and shows the need for trusted repositories across disciplines. It also points to the need for educating scientists about how they can help make their data more easily accessible through good data practices.

Of course having a place to put data is only part of the story. Building in habits of going to trusted sources for data and information is another part. More than half of all respondents in the south east agreed with the statement: knowing where to find information I need is a challenge. Just under half agreed with the statements: "the best way to find information is to ask a coworker or colleague‖ (47%), and –finding information I need is difficult and takes too long‖ (44%). A majority believe the information they need is available, somewhere. Helping scientists improve their information seeking skills will increase the usage of data sources, as search tools were rated as the most important information tool by IBIS respondents. In another question on the IBIS survey, a majority (55%) indicated that they believe the information they need is available, but knowing where to find the information they need is a challenge.

Scientists have stringent requirements for their biodiversity information sources. A vast majority of respondents in the Southeast rated each of the attributes shown in Fig. 4 as important or essential. Trust is number one, followed closely by provenance and completeness. The fact that more than 95% consider that it is important for the information to come from a trusted source suggests that a trusted brand adds value to any resource. It also suggests that it is important for organizations to follow processes that assure the quality of the resource. Navigation and usability were also found to be very important in the IBIS survey (see Fig. 4).

This leads directly to restrictions and conditions for data sharing. As we saw above, most scientists are willing to place at least some of their data into a central repository. Most agreed they would be willing to use other's data sets, share their own data sets, and that it is appropriate to create new datasets from shared data.

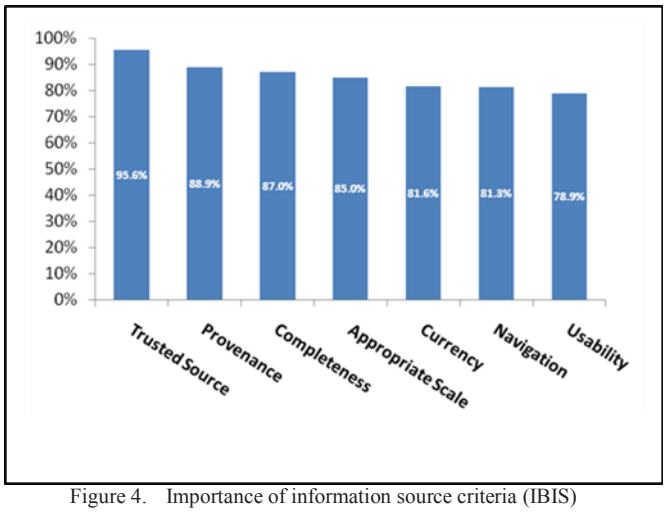


Figure 4. Importance of information source criteria (IBIS)

Researchers may need assurances of security and that their data rights are protected. Scientists in the DataONE survey identified many conditions necessary as conditions for –fair exchange‖ of data. A vast majority agree that it is a fair condition to require formal citation or acknowledgement of any data sets used. A majority also want the opportunity to collaborate, have reciprocal data sharing, or receive reprints of publications or a complete list of products that used their data (see Fig.5).

When scientists don't make their data available electronically, the number one reason is insufficient time (45%), followed by lack of funding (34%). While we can't put more hours in the day, we can develop the tools and products that allow scientists to work more productively in the time they do have. We also cannot give them more money but through building good partnerships and interoperability we can make the money they have go further. Other reasons are—no place to put the data (20%) and lack of standards (20%)—reasons that organizations such as DataONE and USGS can directly address.

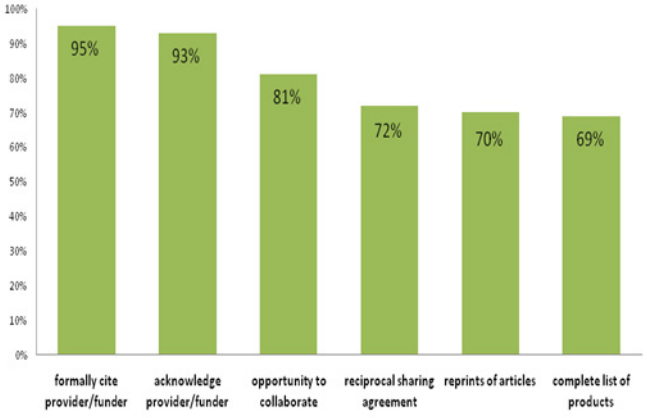


Figure 5. Conditions on data sharing (DataONE).

Government and academic scientists in the interviews agree that there should be certain restrictions and conditions to sharing data. Joe is more concerned with who is using his data, while Mabel wants to be sure that she gets appropriate recognition. Both would likely agree for the necessity of protecting endangered species to restricting access to sensitive data.

Government Scientist Joe:

–We will share it with people who want to use the data for restoration or research. If a consultant wants data to make money, then we are hesitant to hand it out.

–Is there a mechanism by which we can know when our data is being used? Knowing how valuable we are to the general public comes from the use of our data.

Academic Scientist Mabel:

–We want to make sure that those of us who have been involved in gathering the data get appropriate recognition for it.

–If someone were to ask about rare or endangered plants, I would limit that information to appropriate people; natural heritage, universities and federal agencies.

Joe and Mabel reflect common themes from the literature in terms of both data sharing (Lesson 2) and data withholding (Lesson 3), however, by examining government and academic environmental scientists' practices and attitudes separately, it is clear that differences, potentially quite meaningful ones, begin to emerge.

Both Joe and Mabel's enthusiasm and hesitancies about data sharing are indicative of their professional contexts. Mabel's enthusiasm may also reflect the generally positive data sharing culture found within biodiversity research, the sub-discipline from which these interviews were drawn. While Joe must consult the bureaucratic chain of command in order to share his data, Mabel is free to make decisions for herself. However, her hesitancy is indicative of the professional pressures in academia while Joe's indicates the increasing need to be transparent and mindful of the relationship between how public funds are spent and what is gained by the costs. These differences are further discussed in Lesson 4.

Lesson 4: There are different needs, attitudes, and practices who work in academia.

On the whole, academic scientists who responded to our surveys or participated in interviews are much more satisfied with the processes for cataloging or describing their data, and also with tools for documentation. It may be more because they are unaware of metadata standards and practices, however, rather than being satisfied following them. Government scientists in the DataONE survey are much more satisfied with their ability to manage data during the life of their projects, and storing data long-term. (Table 1).

TABLE I. SATISFACTION WITH ABILITY TO ENGAGE IN DATA MANAGEMENT ACTIVITIES (DATAONE)

	% Government	*%Academic*
Satisfied with the process for cataloging/describing my data	47.5	61.5
Satisfied with the tools for preparing my documentation	33.7	45.6
Managing data during the life of the project	52.4	39.6
Storing data beyond the life of the project	53.3	34.6

Government respondents are more likely to agree that their organizations are involved both with short-term data management (that is, during the life of the project) and long-term data management (that is beyond the life of the project), although even then only slightly more than half feel that way. There is much room for organizational leadership including training and policies in both sectors related to data management plans, data description, data deposition and data curation.

Government respondents use several sources more often than do academics. State environmental and wildlife resource agencies are utilized by nearly two-thirds of the government scientists. This suggests that finding ways to facilitate access across these agencies could leverage existing resources and increase use of these critical data. We don't yet know if the lesser usage patterns among academics is a result of a lack of awareness of these sources or if they are less comfortable accessing these sources. The answer to this question will inform how USGS can better reach scientists in the academic community.

Academic respondents are also significantly more likely to have sole responsibility for approving access to some or all of their datasets. This suggests activities to facilitate creating access to these data sets would be successful because these academic scientists have the ability to approve access.

Comments from Joe and Mabel illustrate the different perspectives that reflect being in government and academic organizations. Joe notes that being in government means working within boundaries established by the agency which extends to issues related to data sharing:

–I don't have the authority to make decisions about data sharing.

–Our data sharing policy makes it difficult for us to withhold parts of the datasets we receive. As a result, some data contributors only share sub-sets of their data.

Conversely Mabel has the freedom to establish how she will handle her data but also is highly motivated to be able to cite usage of her data – especially since academics depend on this type of credit for promotion. As Mabel notes:

–I don't have anything I'm keeping private. I'm willing to put it all out there.

And –If other people are using my data then I somehow need to report that. I need to know how it's being used and if any publications result.

Lesson 5: The skill level of scientists and use and access to appropriate tools varies across the data life cycle.

Approximately 40% of scientists in the southeast say biodiversity information is difficult or very difficult to find, yet more than 60% say that half or more than half of the information they need to do their work relates specifically to biodiversity. This suggests several things.

1. Scientists may not have the information seeking skills needed to do their work.
2. The information may not be easily accessible.
3. There is room to improve system navigation and organization.

Scientists in the south east rate a variety of tools as important, with information search right at the top (88%), followed by mapping (81%), data management (68%), visualization (63%), and documentation (63%) tools. This question did not tell us how often these tools are currently used, however the answers suggests how tool development may be prioritized since these were the tools deemed important by scientists.

Although academic scientists are satisfied with the process for cataloging and description of their data, there is little evidence that metadata use is widespread. When we asked scientists internationally what metadata standard they used to describe their datasets, by far the largest choice was –none (56%), followed by their lab's own standard (22%). There is much room for education and training in best practices in use of metadata standards.

Government and academic scientists discussed metadata. Government scientist Joe is clearly working in an environment which values metadata and how that can help not only describe but manage the data. He says:

–For contemporary sets, the person who submits the data also submits a metadata record. We create another record representing what we think it is. We have one version of the data, submitter may have a version they keep on their website. We want to be able to show that these are two different things.

–We write FGDC records.

Conversely, academic scientist Mabel is working in an environment that has little engagement with metadata, although it is being used in natural history collections.

–For my research, very little metadata has been created. For metadata associated with the museum collection, Darwin Core has been used.

There is also activity towards building unique metadata schema rather than adopting standardized ones in the academic community:

–We are currently redoing all of our collection databases at the museum. We are building an in-house system. We looked at available standards and decided to write our own.

Lesson 6: There are many ways that scientists can be assisted across the data life cycle.

Less than half of respondents from government and academia in the DataONE survey feel that their organization currently provides training on best practices or funding or tools for short and long-term data management. Clearly there is an opportunity here for all types of organizations. Even scientists who are willing to share data resources or use those from others need assistance. While it is unlikely there are ways to increase funding, there are opportunities for building partnerships, tools and services that can help facilitate data management in the long and short term – including by providing training on best practices which can improve the efficiency of data producers.

Government and academic scientists' comments suggest how they would like to be assisted.

Joe notes:

–Ideally, we would like for our research results to be disseminated in a way that's accessible and digestible to not just academics but to everybody.

–Manpower. We need more people to handle these sorts of things.

Mabel needs help with geo-referencing and data integration.

–Maximum utility of the data would require geo- referencing of the data. We would need help geo-referencing the part of the collection that isn't geo-referenced.

–It is cumbersome to put those data sets together, but only because it is important. If there were ways to automate some of that information collection out of the data sets, it would help.

III. CONCLUSION

Assessment is a strategic tool to highlight barriers and opportunities of what can be done at each stage of the data life cycle to increase participation in better practices of data management and, ultimately, to help scientists access and use the information and data resources they need to improve science into the future.

Similar to previous studies, the results of these assessments show that environmental scientists are willing to share data, with some restrictions. Adding to the literature, are findings concerning differences in practices and attitudes across scientific work sectors with regards to all aspects of the data life cycle. For all scientists, there are still many challenges to improved data management throughout the data lifecycle. Many scientists need assistance, through education, training, good systems, and access to trusted sources. This presents an

opportunity for libraries, data centers, and data curation and information specialists to assist.

ACKNOWLEDGMENT

Many people in USGS and the DataONE project were involved in this work, notably Mike Frame, Elizabeth Martin, and Jean Freeney of USGS; William Michener and Trisha Cruz of DataONE and members of the Usability & Assessment and Sociocultural Working Groups. Special thanks to University of Tennessee staff and graduate assistants, including Chris Caldwell, Liz Whitson, Jana Redmond, Lei Wu, and Arsev Aydinoglu.

REFERENCES

[1] [1] C. Tenopir, S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame, ―Data sharing by scientists: Practices and perceptions,‖ PLoS ONE, vol. 6, Iss. 6, June 2011.

[2] [2] PARSE Insight, ―PARSE Insight, December 2009, retrieved from http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

[3] [3] Publishing Research Consortium, *Access vs. Importance,* October 2010. Retrieved from http://www.publishingresearch.net/documents/PRCAccessvsImportanceGlobalNov2010_000.pdf

[4] [4] P. Arzberger, P. Schroeder, A. Beaulieu, G. Bowker, K. Casey, L. Laaksonen, D. Morroman, P. Uhlir, and P. Wouters, ―*Promoting access to public research data for scientific, economic, and social development*,‖ Data Science Journal, vol. 3, pp. 135-153 , November 2004.

[5] [5] C. J. Savage and A. J. Vickers, ―*Empirical study of data sharing by authors publishing in PLoS journals*,‖ PLoS ONE, Vol. 4. e7078.doi: 10.1371/journal.pone.0007078, September 2009.

[6] [6] E. G. Campbell, B. R. Clarridge, M. Gokhale, L. Birenbaum, S. Hilgartner, N. A. Holtzman, and D. Blumenthal, ―*Data withholding in academic genetics: evidence from a national survey*,‖ The Journal of the American Medical Association, vol. 287, pp. 473-80, January 2002.

[7] [7] USGS Strategic Directions Statement

[8] [8] B Meko. 2010. *Biodiversity Information Needs of Environmental Decision Makers: A Literature Review.* Report for USGS via the IBIS project. University of Tennessee, Knoxville.

[9] [9] Campbell EG, Bendavid, E (2003) Data-sharing and data-withholding in genetics and the life sciences: Results of a national survey of technology transfer offices. J of Health Care Law Policy 6: 241-255

The Initial Design of Data Sharing Infrastructure for the Critical Zone Observatory

Ilya Zaslavsky[1], Thomas Whitenack[1], Mark Williams[2], David Tarboton[3], Kim Schreuders[3], Anthony Aufdenkampe[4]
[1] San Diego Supercomputer Center
[2] University of Colorado at Boulder
[3] Utah State University
[4] Stroud Water Research Center
zaslavsk@sdsc.edu, twhitenack@sdsc.edu, markw@cutler.colorado.edu, dtarb@usu.edu, kim.schreudres@usu.edu, aufdenkampe@stroudcenter.org

***Abstract*—The Critical Zone Observatory (CZO) program is a multi-institutional collaborative effort to advance scientific understanding of environmental interactions from bedrock to the atmospheric boundary layer across scales and disciplines. To create a comprehensive hydrogeochemical portrait of experimental sites the observatories collect large volumes of data. Publishing, analyzing and archiving these data in a consistent and integrated manner across all CZO sites is challenging due to the inherent heterogeneity in data collection and processing techniques. We present the initial design and a prototype of the CZO data sharing infrastructure. While each CZO site maintains its own data management system, the integrated infrastructure design specifies formats and protocols for presenting the information on CZO web sites, where it can be browsed by users as well as automatically harvested into a centralized data system. The latter validates, archives and converts the data into standards-compliant data services, which can be consumed by various client applications.**

Keywords—environmental observatory; CZO; cyberinfrastructure; hydrology; information integration

I. INTRODUCTION

The CZO project [1] integrates data from several earth science disciplines in order to describe and model complex physical processes in the critical zone. Typical research scenarios involve accessing both geochemical samples and hydrologic time series of water quality and water quantity within experimental watersheds, relating the dynamics of differently measured parameters, modeling soil nutrients under different topographic, geologic, hydrologic and vegetation conditions, analysis of fluxes across watershed boundaries, etc. While closely connected research teams have been successful in such cross-discipline analysis and modeling, accomplishing such integration at a higher level, across CZO sites and spatio-temporal scales, faces several interoperability challenges. They stem, in particular, from differences in information models used in different disciplines and by different research groups to describe observations, differences in data representation and access, and discrepancies in metadata and their semantics. For example, the geochemical community has been developing infrastructure for managing geochemical sample information and created a standard XML schema encoding for geochemical datasets named EarthChem XML [2, 3]. The hydrologic research community, via the Consortium of Universities for the Advancement of Hydrologic Science, Inc.'s Hydrologic Information System (CUAHSI HIS) project, has been creating a service-oriented system for sharing hydrologic observations [4, 5], and proposed a canonical data model for hydrologic observations [6] encoded as Water Markup Language [7]. Large scale cross-observatory systems are being developed within the Long Term Ecological Research Network [8], the National Ecological Observatory Network [9], and several other NSF-supported earth science projects. Common cyberinfrastructure challenges of earth science observatory projects have been summarized in [10].

The CZO program is a relatively new large-scale observatory effort, which allows the CZO information network design to leverage the experience and cyberinfrastructure components developed in the neighboring projects. It currently includes 6 observatories: the Boulder Creek CZO (led by the University of Colorado at Boulder), the Christina River Basin CZO (University of Delaware), the Jemez River and Santa Catalina Mountains CZO (University of Arizona), Luquillo CZO (University of Pennsylvania), the Southern Sierra CZO (University of California, Merced) and the Susquehanna Shale Hills CZO (Pennsylvania State University). Research agendas of each site are different, yet several cross-cutting topics and data needs have been identified, in particular with management of hydrologic time series; water, soil and rock samples; spatial data including LiDAR; and meteorological variables. This justifies development of a CZO-wide data sharing infrastructure, to enable uniform publication, discovery and retrieval of data collected across sites.

Despite differences in research foci and scope, the experience of large environmental observatory cyberinfrastructure projects suggests multiple common requirements and infrastructure issues; they have been addressed in the literature [e.g. 10, 11, 12, 13]. Specific requirements of the CZO-wide data management system derive from the unique role of the CZO program as an evolving cross-disciplinary multi-site effort. They can be summarized as follows:

- Reliance on standards for data exchange adopted in research communities comprising the CZO program.

- Leveraging domain data systems, synthesizing information management experience and software from CZO partners and neighboring earth science disciplines (CUAHSI [4, 5], EarthChem [2, 3], CZEN [14], NCED [15], LTER [8], etc.)
- Uniform data modeling, data description and formatting practices, to ensure that the published data can be unambiguously interpreted and their provenance can be traced.
- CZO research teams maintain their own data management systems, while sharing data via a centralized publication system that is scalable and extensible to additional data types and research sites.
- Evolving the integrated data system towards better standards compliance and cross-CZO integration without burdening individual CZO sites.
- Availability of CZO data both in a human-readable form at individual CZO web sites as well as via web services from the central CZO data repository.

This paper presents the details of the original design of the CZO-wide data publication and sharing system developed in response to these requirements, and describes its main components.

II. The Vision of the CZO Data System

The CZO project is enabling access to a variety of data types required for modeling physical processes in the critical zone, including geochemical, geophysical and hydrologic observations, spatial data and field measurements. For some types of data, uniform protocols and formats for data and metadata exchange have been established and agreed upon within respective communities, while other domains see a wide variety of approaches to data representation and description. Therefore, we consider CZO data interoperability at several levels (Fig. 1).

At the first level, different types of CZO resources (files, services, downloadable data folders, etc.) are registered at a CZO data portal, with Dublin Core metadata, so that these resources can be browsed or queried by title, contributor, spatial location, thematic category and similar fields as defined in the Dublin core standard, and subsequently invoked or downloaded to a user's workstation.

At the next level, the resources have common semantics (a set of shared vocabularies for variable names, methods, units, features of interest, measurement medium, qualifiers, censor codes, etc.) which ensures that, once the resources are discovered and downloaded they could be easier interpreted and integrated.

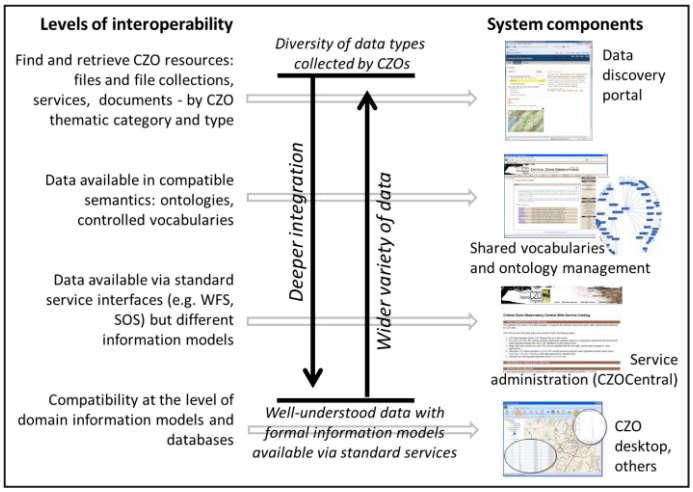


Figure 1. Levels of interoperability and corresponding components of the CZO data system

Further, resources of certain types may become available via standard service interfaces, such as those developed by the Open Geospatial Consortium (OGC), so that they can be accessed from standards-aware client applications.

Finally, at the fourth level the data become available via standard services and in standard encodings that reflect domain information model, to enable a much wider range of operations across different compliant sources.

Different types of data considered by CZO support different levels of interoperability, and, therefore, rely on different system components. For example, soil samples, gridded data, flux information are currently registered as resources with minimal metadata and made available via the data discovery portal, while their semantic consistency is recommended by a set of shared vocabularies but is not currently enforced, and standard information models are being developed. Hydrologic observations, on the other hand, represent the type of data that is made interoperable at all four levels within the CUAHSI HIS project. In the current design, the CZO data infrastructure leverages HIS components and generally follows Service Oriented Architecture (SOA) for publishing, indexing and accessing hydrologic observations as implemented in the HIS project.

The CZO data system design follows the general SOA "publish-find-bind" pattern, with the additional requirements described above. In particular, these requirements affect the "publish" component which is represented as two interlinked modules: publishing CZO data at individual web sites as human-readable ASCII files, and their harvesting and republishing as standard-compliant web services at the central CZO data repository/archival site (Fig. 2). The system components supporting CZO data interoperability at different levels are described in the following section.

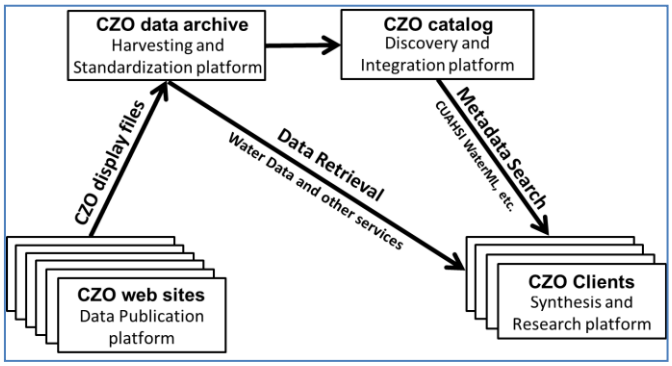


Figure 2. A high-level view of service oriented architecture patterns in the CZO data publication and sharing system

III. THE PROTOTYPE: MAIN COMPONENTS AND INTERFACES

This overall design is further detailed in Fig. 3. Data published at each of the six CZO web sites following an agreed upon ASCII format (display file format, described below), are automatically harvested into a centralized data repository housed at the San Diego Supercomputer Center (SDSC), validated against shared vocabularies and parameter ontology and archived in a set of databases established for each CZO. Upon harvest and validation, standard CZO data services are automatically updated to include the new data. The CZO data services become available in a range of standard formats: for hydrologic observations these are CUAHSI WaterOneFlow services, which transmit data according to the WaterML 1.x specification, and Web Feature Services (WFS) specified by the Open Geospatial Consortium, which are used to exchange time series catalog information. The services are registered and indexed in the CZO Central's service registry, and can be discovered via a CZO Data Portal, which is compliant with OGC's Catalog Services for the Web (CSW) standard. The standard services can then be consumed by various applications, as well as registered in cross-project domain registries such as CUAHSI HIS Central (for hydrologic time

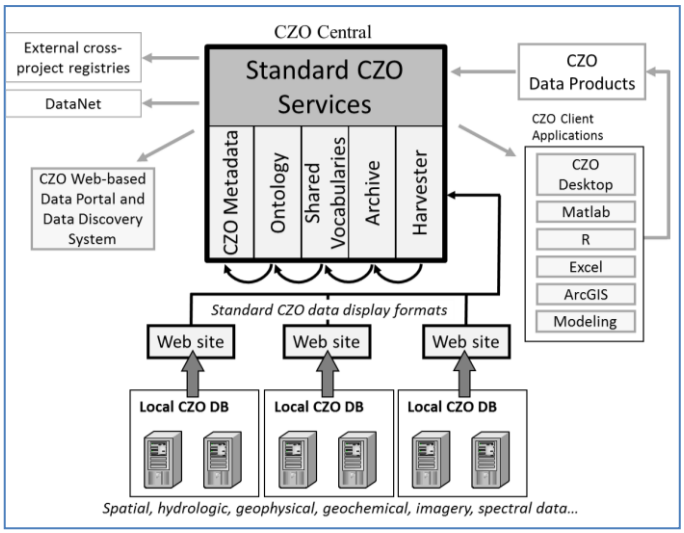


Figure 3. Main components of the CZO data publication workflow

series) or the EarthChem Data Portal (for geochemical data). CZO data products generated from the published data will be made available via the same CZO Central services and portal. In addition, we envision collaboration with the NSF-supported DataNet program on long-term preservation of CZO data.

Below we describe several key components of this design, which is now implemented in a prototype data sharing system.

A. CZO Display File Format

The CZO display file format for hydrologic time series has the following key features:

- The format is based on the information model adopted from CUAHSI Observations Data Model (ODM) [6]. At the same time, it incorporates several information model enhancements made necessary by CZO data collection practices, including multiple types of named vertical offsets (e.g. "upper canopy", "lower canopy"), support for data loggers collecting information from groups of sampling locations, and a more flexible definition of a data series as any logical grouping of observations defined by data publisher.
- The ASCII format of the display files is both human- and computer-readable, and is uniform across the CZO sites.
- Display files include a configuration file (specifying which files shall be regularly harvested from a CZO web site), sites file, methods file, series metadata file, and a data file. In a typical scenario, each configuration file housed by a CZO will point to single sites and methods files, and to one or more series metadata files, each of which would reference one or more data files.
- The display data file closely follows a common data logger format, to minimize re-formatting at CZOs. It encodes the following characteristics of each observed value: location (where the observation took place), date and time (when), the attribute measured (what), the measurement method (how), and the responsible investigator (who). Details of each of these characteristics are encoded in the sites, methods and header (series metadata) files.
- While initially focused on hydrologic time series, the display file format is extensible to other types of data, in particular geochemical samples. Further, metadata display files (configuration, sites, methods, data series) may reference binary data files if appropriate for certain data types (e.g. spatial data, grids).

B. CZO Central Catalog and Web Services

The CZO Central model generally follows the organization of the centralized components of the CUAHSI Hydrologic Information System [16] and extends it to accommodate the specific CZO data management requirements: managing centralized rather than distributed collection of ODM databases and supporting harvesting and validation of display files. New or updated display files are being retrieved from each of the six CZO web sites into the CZO central data repository (currently

configured to re-harvest new data automatically every week or manually by request from a data manager). The harvesting triggers updates of respective ODM instances installed at the CZO Central for each site, along with validation of the display files configured by CZO data managers. Through the CZO Central's online interface, data managers can browse harvesting logs and correct errors if necessary. Once the central ODM instances are updated, the time series metadata are harvested into the CZO Central time series catalog, which makes the data from all CZO sites discoverable by a range of spatial, temporal and semantics-based requests.

The data in each ODM are available via a standard set of water data services compliant with the WaterML 1.x specification [7]. For each CZO hydrologic observation network the services include the following standard methods: GetSites, GetSiteInfo, GetVariableInfo and GetValues. Once harvested into the central catalog, metadata from all CZO sites become available via requests that return time series information based on spatial, temporal and attribute-based requests (GetSeriesCatalogForBox), site information (GetSitesInBox), information about services (GetServicesInBox, GetWaterOneFlowServiceInfo), as well as information about searchable concepts and their hierarchy (GetOntologyTree, GetSearchableConcepts, GetWordList), and mapping between variables and concepts (GetMappedVariables). Compatibility with CUAHSI HIS at the level of services and information models makes it easy to integrate CZO data with data available from over 70 government and academic observation networks available through CUAHSI HIS Central. This enables easier validation of CZO-collected data against hydrologic observations made at USGS, EPA, and possibly collocated stations from other networks, and additional data interpolation/imputation processing.

CZO data managers login to the CZO Central administration interface to edit service metadata for their sites: abstract, contact information, recommended citation, data access policy, icons/logos, etc. (Fig. 4), request harvest of their published display files into the central system, and examine the harvesting logs. In addition, data managers can use the CZO Central application to associate variable names with concepts in the ontology of hydrologic terms developed by CUAHSI. Establishing this association enables data discovery based on thematic categories. The CZO Central web site is central.criticalzone.org.

C. *CZO Data Portal, and compliance with OGC services*

Besides making the time series metadata available via WaterOneFlow and CZO Central web services, the CZO Central application also generates WFS services for each CZO network, which list time series available from each site, and their metadata. These services are automatically registered in the CZO Data Portal, a custom application of the ESRI open source GeoPortal Server [17]. With this application, the registry of CZO services becomes available via standard OGC Catalog Services for the Web (CSW) interface, which makes them accessible from a variety of OGC-compatible client applications, and enables federation with other CSW catalogs, such as the CUAHSI HydroCatalog. Sample search results in

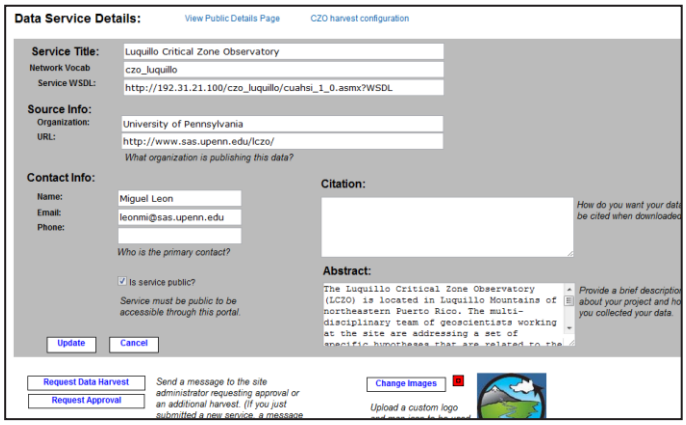


Figure 4. A fragment of a CZO service management and metadata editing web page at CZO Central

the CZO Data Catalog, federated with the HydroCatalog at CUAHSI, are shown in Fig. 5.

In addition to registering water data services to the CZO Data Portal, the harvesting application automatically adds display files retrieved from CZO web sites, to the same central CSW catalog, thus enabling full text search over the content of registered metadata files, and data file download directly from the portal application.

One of the key roles of the CZO Central and the CZO Data Portal is to expose CZO data via standard OGC-compliant service interfaces, and evolve these interfaces once new specifications are adopted. With respect to hydrologic data, an essential new standard is WaterML 2.0, which is being developed under the aegis of the Hydrology Domain Working Group of the OGC and the World Meteorological Organization [18]. At the time of writing, this specification, after being

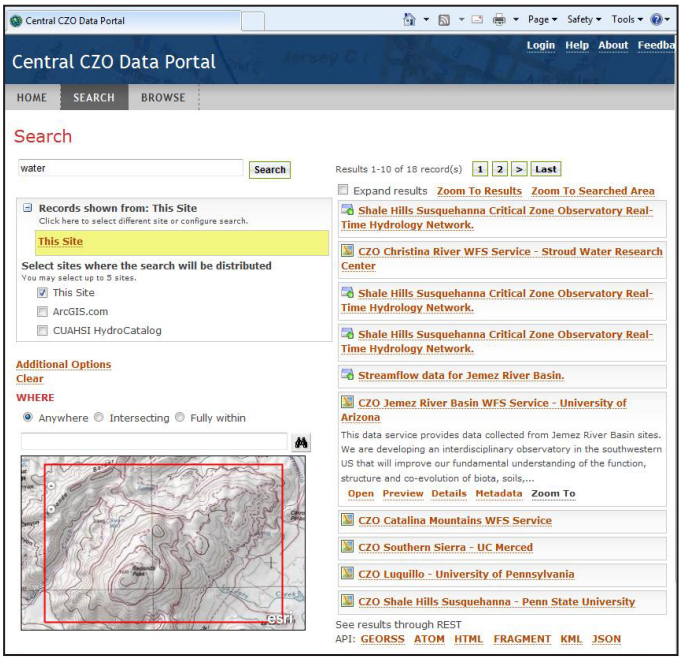


Figure 5. The search page of the CZO data discovery portal

approbated and refined through OGC Interoperability Experiments, is entering the OGC standardization process, and is expected to be approved by the end of 2011. The first version of WaterML 2.0 is a profile of the OGC/ISO "Observations & Measurements" [19] model and specifies time series encoding for hydrologic data. Thus encoded time series data will be transmitted over OGC Sensor Observation Service (SOS) 2.0 interface, initially alongside WaterML 1.x/WaterOneFlow services, and eventually replacing them for both CUAHSI HIS and the CZO data system.

D. CZO Shared Vocabulary Registry

Another key component of the CZO data system is a collection of controlled (shared) vocabularies, also inherited from the CUAHSI HIS ODM controlled vocabulary submission system [20] but extensible to other types of data collected by CZO sites. These vocabularies, available via a web interface and via web services, are used to establish semantic conventions within the CZO system, ensuring that terms describing key metadata elements are well defined, unique and unambiguous, which, in turn, supports cross-CZO attribute-based data discovery. The web interface for the shared vocabulary system allows data managers to browse the vocabulary content, and propose additions and edits, while the web service API is used by the CZO Central's harvesting application to validate submitted metadata for compliance with the vocabularies. The following vocabularies are moderated by the system: variable names; methods; units; value type (e.g. field observation, model output); sample type (physical medium from which the sample is taken); data type (e.g. average, continuous, cumulative); data level (processing level or quality control level); spatial reference (projection and datum, based on EPSG [21]); censor code (e.g. not-censored, non detect); qualifier code (e.g. approved, provisional); vertical datum. If a particular term is missing from any of the vocabularies, data managers can submit it via the web interface; once the term is considered and accepted by vocabulary curators it becomes part of the master list of approved vocabulary terms. The web site for the CZO shared vocabulary registry is sv.criticalzone.org.

IV. DISCUSSION AND CONCLUSION

The initial effort to design and build an integrated CZO data system prototype has achieved several important goals: the CZO sites have converged on a uniform data publication model and a display file format convention, enhancements to the original information model for hydrologic observations have been developed and tested, the initial centralized data system has been built to share and integrate data from all CZO sites, and each CZO has started publishing the data through the system. Most importantly, the system has been designed and developed in close collaboration with data managers from all CZO sites, taking into account differences in data types, metadata organization and data publishing practices established at each site.

While following the CUAHSI HIS architecture, the CZO data system presents a new publication model, which reflects specific requirements of the CZO cross-site and cross-domain data integration. The key advantages of this model include:

- Individual CZO sites are responsible for maintaining their own data systems and are not required to install and maintain additional software (e.g. a HydroServer, which represents the data publication platform within CUAHSI HIS), which may not fit with the existing software environment or skill set of local data management personnel. Developing an ASCII export into the display file format usually presents a lesser problem compared with the need to manage additional software.
- The data publication and sharing model preserves the autonomy of individual research sites, which reflects the level of autonomy of investigator teams in this large and complex project, and thus does not violate the established relationships and practices of the CZO virtual organization.
- The burden of compliance with evolving standard service interfaces and encodings is on the central data management system, rather than on individual CZO sites, where research and data management work can remain focused on science objectives of each site.
- The developed display file format serves a dual purpose: it presents the data in a human-readable form on CZO web site, and at the same time supports automatic harvesting of the data into the central data repository.
- The publication model is extensible to other types of data (raster data, GIS layers, geochemical data, soil profiles, geophysical data, photos, etc.), once respective information models and metadata profiles are agreed upon.

At the same time, these advantages underscore the core drawback of the publication model: it introduced a new exchange format, which needs to be governed and further developed as CZO needs evolve – rather than passing the governance burden to standards organizations such as OGC. Being a text format, it provides limited options for content validation of the display files – which is to some extent compensated by extensive content validation as the files are harvested into the CZO Central repository.

The described CZO information system prototype creates new opportunities for critical zone environmental observatories to publish and discover data and integrate them in new types of cross-CZO data-intensive analysis and modeling that were not possible or too time-consuming before. While the system is at an early development stage (at the time of writing, only about 15 million hydrologic observations collected by CZO sites are available via web services, and about 70 resources are registered in the CZO Data Portal), the volume of data is growing. The prototype demonstrated that a scalable data sharing infrastructure for environmental observatories can be built by leveraging and integrating service-oriented approaches and cyberinfrastructure components developed in neighboring Earth science disciplines, while careful consideration is given to the specific requirements of the CZO research community, in particular: information modeling needs; standards compliance and semantic consistency; and distribution of data

management roles and responsibilities between individual sites and the central archival, cataloguing, and services system.

ACKNOWLEDGMENT

We are grateful to data managers from the six CZO sites for their contribution to the development of the integrated CZO data system and for useful discussions: Matej Durcik (University of Arizona), Chi Yang (University of Colorado at Boulder), Otto Alvarez and Xiande Meng (University of California, Merced), Miguel Leon (University of Pennsylvania), Brian Bills and Jennifer Williams (Pennsylvania State University), Charles Dow and Melanie Arnold (Stroud Water Research Center). U.S. National Science Foundation support (awards EAR-0724960, DEB-1027341) is gratefully acknowledged. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] National Critical Zone Observatory (CZO) Program, www.criticalzone.org (last accessed June 1, 2011).

[2] The EarthChem Portal, www.earthchem.org (last accessed June 1, 2011).

[3] K. Lehnert and S. Vinayagamoorthy, "Geoinformatics for Geochemistry (GFG): Integrated digital data collections for the earth and ocean sciences", Proceedings of Geoinformatics 2007—Data to Knowledge. USGS Scientific Investigation Report 2007-5199, U.S. Geological Survey, Reston, VA, pp. 32-34, 2007.

[4] D. G. Tarboton, J. S. Horsburgh, D. R. Maidment, T. Whiteaker, I. Zaslavsky, M. Piasecki, J. Goodall, D. Valentine and T. Whitenack, "Development of a community hydrologic information system," 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation, ed. R. S. Anderssen, R. D. Braddock and L. T. H. Newham, Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation, July 2009, p. 988-994.

[5] D. G. Tarboton, D. R. Maidment, I. Zaslavsky, D. P. Ames, J. Goodall, and J. S. Horsburgh, "CUAHSI Hydrologic Information System 2010 Status Report," Consortium of Universities for the Advancement of Hydrologic Science, Inc, 2010, 34 pp., http://his.cuahsi.org/documents/CUAHSIHIS2010StatusReport.pdf (last accessed June 1, 2011).

[6] J. S. Horsburgh, D. G. Tarboton, D. R. Maidment and I. Zaslavsky, "A Relational Model for Environmental and Water Resources Data," Water Resour. Res., 44: W05406, 2008.

[7] I. Zaslavsky, D. Valentine and T. Whiteaker, "CUAHSI WaterML," OGC 07-041r1, Open Geospatial Consortium Discussion Paper, 2007, http://portal.opengeospatial.org/files/?artifact_id=21743 (last accessed June 1, 2011).

[8] The US Long Term Ecological Research Network, www.lternet.edu (last accessed June 1, 2011).

[9] National Ecological Observatory Network, www.neoninc.org (last accessed June 1, 2011).

[10] Cyberinfrastructure for Environmental Observation Networks (CEON) Workshop Report, held February 25 & 26, 2008 at The National Science Foundation, Arlington, VA. http://feon.wdfiles.com/local--files/start/Feb2008WorkshopFinalReport.pdf (last accessed June 1, 2011).

[11] J. Dozier and W. B. Gail, "The emerging science of environmental applications", In "The Fourth Paradigm", T. Hey, S. Tansley and K. Tolle, Eds., Microsoft Research, Redmond, WA, pp. 13-29.

[12] M. Lehning, N. Dawes, M. Bavay, M. Parlange, S. Nath and F. Zhao, "Instrumenting the earth: next-generation sensor networks and environmental science", In "The Fourth Paradigm", T. Hey, S. Tansley and K. Tolle, Eds., Microsoft Research, Redmond, WA, pp. 45-51.

[13] J. S. Horsburgh, D. G. Tarboton, D. R. Maidment and I. Zaslavsky, "Components of an environmental observatory information system," Computers & Geosciences, 37(2): 207-218, 2011.

[14] Critical Zone Exploration Network, www.czen.org (last accessed June 1, 2011).

[15] National Center for Earth Surface Dynamics, University of Minnesota, http://www.nced.umn.edu/ (last accessed June 1, 2011).

[16] T. Whitenack, I. Zaslavsky, D. W. Valentine, "HIS Central and the hydrologic metadata catalog", Eos Trans. AGU, 89(53), Fall Meet. Suppl., Abstract IN51A-1142, 2008.

[17] ESRI Geoportal Server, http://geoportal.sourceforge.net/ (last accessed June 1, 2011)

[18] Open Geospatial Consortium – Hydrology Domain Working Group, http://external.opengis.org/twiki_public/bin/view/HydrologyDWG/ (last accessed June 1, 2011).

[19] Observations and Measurements – XML Implementation, OGC Document 10-025r1, http://portal.opengeospatial.org/files/41510 (last accessed June 1, 2011).

[20] J. S. Horsburgh, D. G. Tarboton, M. Piasecki, D. R. Maidment, I. Zaslavsky, D. Valentine and T. Whitenack, "An integrated system for publishing environmental observations data," Environmental Modelling & Software, 24(8), pp. 879-888, 2009.

[21] European Petroleum Survey Group, Geodetic Parameter Dataset, http://www.epsg-registry.org/ (last accessed June 1, 2011).

A Semantically-Enabled Provenance-Aware Water Quality Portal

Jin Guang Zheng[1], Ping Wang[1], Evan W. Patton[1], Timothy Lebo[1], Joanne S. Luciano[1] Deborah L. McGuinness[1]
[1] Tetherless World Constellation, Rensselaer Polytechnic Institute
{zhengj3, wangp5, pattoe, lebot, jluciano, dlm}@cs.rpi.edu

Abstract— **Environmental informatics systems often analyze data collected from various sources. Both data collection and data analysis benefit from expert knowledge. However, if these applications are to be used by a broader range of users with less expert knowledge, applications will need to include a deeper understanding of the data used and analysis performed. We present the *Tetherless World Constellation Semantic Water Quality Portal* as both a water quality portal application and as an example of a semantic approach to environmental informatics applications. The portal integrates water data from multiple sources and captures the semantics of the data in a simple water quality ontology. Portal users can identify polluted water sources and polluting facilities according to multiple regulation perspectives and geographic constraints by using visualizations of semantically-enabled queries. Further, knowledge provenance is encoded in the data capture and integration services to enable provenance-based queries and reasoning capability.**

Keywords—Semantic Web; Visualization; Semantic Environmental Informatics; Water Quality Portal

I. INTRODUCTION

Water quality has been a major concern for environmental scientists and local citizens who understand the important role that clean water plays in our lives and the health of our planet. Polluted water sources, the kinds of pollutants, and those responsible for the pollution need to be discovered so that corrective and preventative measures can be undertaken. To monitor and control water quality, government agencies such as the Environmental Protection Agency (EPA [1]), U.S. Geological Survey (USGS[2]), regularly collect water quality data about pollutants and establish regulations to define pollution in terms of acceptable levels of pollutants.

To enable citizens and professionals to better utilize these data, environmental informatics systems need to automatically integrate and analyze the data. Such need is reflected in our motivating example in 2009, in Bristol County, Rhode Island, where the cause of diarrhea in children was found to be polluted water. Local citizens expressed concerns such as "when did the contamination begin?", "how did this happen?", and "how well-equipped is the BCWA to monitor and prevent future events?"

However, informatics systems for water quality investigation face the following challenges. 1) Raw data from multiple sources are stored in different formats, e.g. CSV, HTML, TXT, which makes it difficult to integrate and query the data. In addition, the semantics of the raw data are often not machine-accessible, i.e. they cannot be handled by a computer program. 2) The semantics of the water quality data are not explicitly encoded in the data files, but are instead described in help pages on web sites, although not in a machine-understandable format. 3) Analysis over the collected data are often time consuming, since data can be large due to large spatial regions or long time spans. 4) Some of the analysis tasks can be complex. For example, to identify if a water source is polluted, we need to compare all measurements of all pollutants with their corresponding limits in the adopted water regulations.

In this paper, we present the Tetherless World Constellation Semantic Water Quality Portal (TWC-SWQP). The portal is used to detect water pollution. Here, *water pollution* refers to those situations where the measured concentration of one or more characteristics in water samples exceeds numeric criteria for drinking water sources. The portal can identify point sources of water pollution, including water sites monitored by USGS and polluting facilities regulated by EPA. The portal also demonstrates the effectiveness of semantic web technologies in addressing the challenges faced by environmental informatics systems. We designed ontologies to model the domain of water quality investigation and explicitly encode the semantics of the data. Then, data from different sources were converted into RDF triples compatible with the ontologies. In this way, we achieved consistent and machine accessible semantics for the converted data. We load the data into a triple store and retrieve data required by users' queries with SPARQL. Furthermore, we reason over the retrieved data to detect water pollution with a semantic reasoner. In the remainder of this paper, we describe our design and implementation, highlight the benefits of our semantic approach, and discuss the potential impact of this approach for water quality informatics systems and other similar informatics needs.

II. METHODS

A. SWQP System Architecture and Components

The system architecture of the TWC SWQP is illustrated in Fig. 1. The system comprises six major components: (a)

[1] http://www.epa.gov/
[2] http://www.usgs.gov/

ontology, (b) data conversion, (c) storage, (d) reasoning, (e) visualization and (f) provenance.

Ontology Component: There are two types of ontologies in the SWQP: the core ontology and the regulation ontology. The core TWC Water ontology[3] consists of 18 classes, 4 object properties, and 10 data properties. It extends existing best practice ontologies, including SWEET [3] and OWL-Time [4]. The core ontology models domain objects (e.g. water sources, facilities, measurements, and pollutants) as classes, and includes terms for relevant pollution concepts. For example, a polluted water source is modeled as the intersection of water source and something that has a pollutant measurement outside of an allowable a range. The application can use the core ontology to conclude "any water source that has a measurement outside of its allowable range" is a polluted water source. Further, it can discover pollution with respect to any particular pollutant such as arsenic. Subsequently, we can identify a polluted water source with respect to a particular pollutant, given an existing constraint. For example, the portal can identify water sources that are polluted with arsenic, given the rule that arsenic concentrations value greater than 0.01 mg/L may cause adverse health effects.

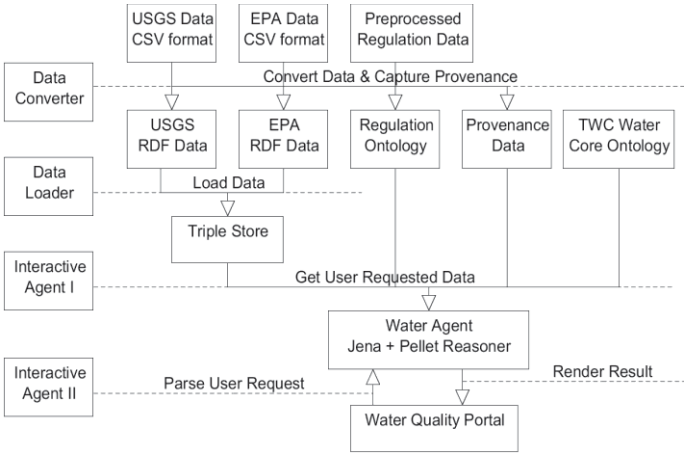


Figure 1. SWQP System Architecture and Workflow[4]

The regulation ontologies[5] model the federal and state water quality regulations for drinking water sources. For example, in California, the state regulation defines 0.01 mg/L as the limit for Arsenic. Because regions differ in their ecology and each state is responsible for its own regulations, the number of pollution concepts (pollutants and limits) and properties vary.

Portions of the core TWC Water ontology and Regulation Ontologies are illustrated in Fig. 2 and Fig. 3.

Data Conversion Component: We use two software tools to convert data into Resource Description Framework (RDF) [5] representations: the open source tool csv2rdf4lod[6] and an ad-hoc converter we developed for SWQP. The general-purpose csv2rdf4lod tool converts tabular data into well-structured RDF according to declarative parameters encoded in RDF [6]. To convert SWQP data, we wrote several conversion parameters to map properties of the raw data to those in our ontologies. One advantage of using the csv2rdf4lod tool is the provenance it captures as we convert the data, which we discuss below.

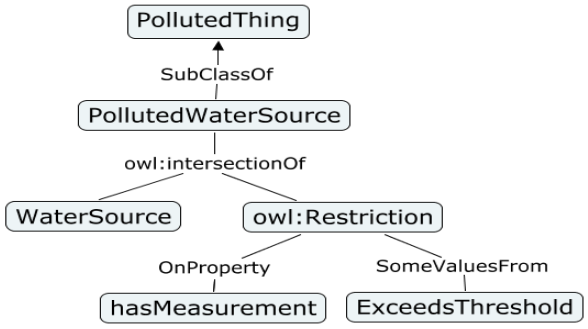


Figure 2. Portion of the TWC Water Ontology.

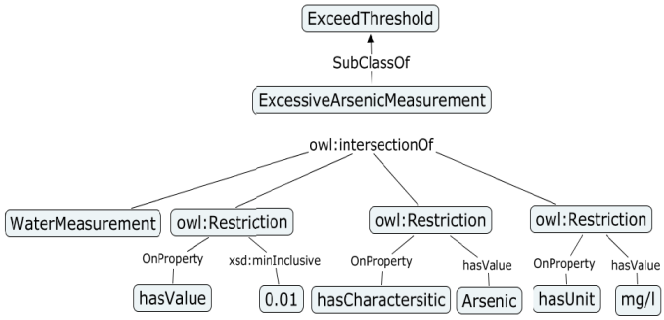


Figure 3. Portion of EPA Regulation Ontology.

To construct OWL 2 [7] constraints that align with rules and properties in our TWC water ontology, we wrote ad hoc converters to extract regulation data from HTML web pages.

Storage Component: Data and ontologies supporting the SWQP were stored in OpenLink's Virtuoso 6[7] open source community edition triple store, which includes a web-accessible SPARQL [8] endpoint[8] that answers queries from web clients.

Reasoning Component: We utilize the Pellet OWL Reasoner [9] together with the Jena Semantic Web Framework [10] to reason over the data and ontologies in order to identify polluting facilities and polluted water sources. Using the core ontology, we model water quality determinations such as "any water source that has a measurement that exceeds a regulation threshold, is to be considered a polluted water source"; using the regulation ontology, we model regulation criteria data, which are region-specific, e.g. California water regulation stipulates: "the threshold for Arsenic is 0.01 mg/L". Combining the above two statements, the reasoning component asserts that any water source that has a concentrattion of aresenic greater than 0.01 mg/L is a polluted water source. At run time, the reasoning component invokes Jena to load the water quality

[3] http://purl.org/twc/ontology/swqp/core
[4] http://was.tw.rpi.edu/swqp/system.png
[5] e.g., http://purl.org/twc/ontology/swqp/region/ny and http://purl.org/twc/ontology/swqp/region/ri; others are listed at http://purl.org/twc/ontology/swqp/region/
[6] http://purl.org/twc/id/software/csv2rdf4lod
[7] http://virtuoso.openlinksw.com/dataspace/dav/wiki/Main/VOSIntro
[8] http://sparql.tw.rpi.edu/virtuoso/sparql

data, the regulation ontology, and the core ontology. Then, Pellet is invoked to classify water sources as polluted or unpolluted from measurements from water samples and their water sources using the regulations as the criteria. The results of this operation can then be queried and visualized.

Visualization Component: This component is responsible for mashing up and representing the data collected from various sources. We support two types of visualizations: (1) map visualization that displays the sources of the water pollution in the context of geographic regions and (2) time series visualization that depicts pollution levels over time with respect to a particular water source or facility.

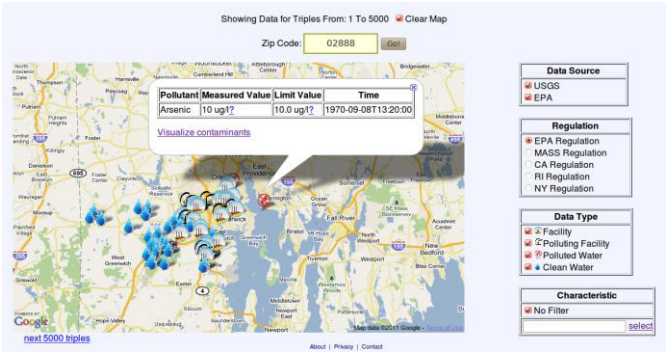


Figure 4. Map Visualization. The results of applying the EPA federal water regulations on the region with zip code 02888 is visualized on a Google Map[9].

The map visualization gets the reasoning results for a user query from the back-end reasoner and displays the results on a Google Map. Different icons are used to distinguish between clean and polluted water sources, and between clean and polluting facilities. Fig. 4 shows an example map visualization. The user may select the data sources to be queried, the regulations to apply, or the types of water sites and pollutants he or she finds interesting. The results of applying the EPA federal water regulation on the region with the zip code 02888 (Warwick, RI) is visualized in this example. Two polluted water sources and eight polluting facilities are indicated with icons.

The time series visualization retrieves water quality data related to a selected water site or facility by querying the triple store and displays the water quality data as a time series using the Protovis visualization toolkit. Fig. 5 shows the phosphorus measurements from 2007 to 2010 in green and the regulation defined limit in blue. Note that the data show one violation in 2009 (in red) and no subsequent violations.

Provenance Component: The portal preserves provenance in the Proof Markup Language (PML) [11] while downloading data, converting data, and loading data to triple store via the provenance support from csv2rdf4lod. The captured provenance data include data sources, the agent that processed the data (i.e. downloaded/converted/loaded), and when the data was processed.

Data source level provenance: The captured provenance data are used to support provenance-based queries. For example, the portal queries the provenance about data sources to get the source organizations for the data and generates the Data Source facet in the map visualization (see Fig. 4). With this facet, the user can select the data organizations he/she trusts and the portal will use only data from the selected organizations.

Reasoning level provenance: When the user clicks a polluted water source or polluting facility in the map visualization (see Fig. 4), SWQP provides explanations in a pop up window for why a water source is marked as polluted or a facility is marked as polluting. The explanations include the names of pollutants, the measured values, the limit values, and the water measurement time. By clicking on the question marks in the pop up window, the user can access supporting provenance data for the explanations including the URLs of the source data, intermediate data and the converted data.

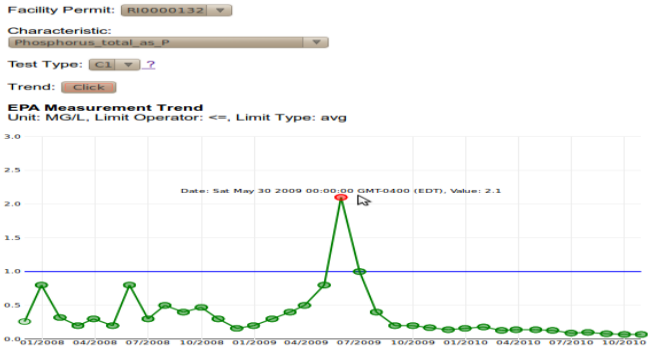


Figure 5. Time Series Visualization. The phosphorus measurements from 2007 to 2010 and the regulation defined limit for the selected facility are visualized[10].

B. *System Workflow*

We now present how the components described in the previous sections work together to identify polluted water sources, polluting facilities, and pollutants. Fig. 1 also shows the system workflow. SWQP first downloads data from USGS, EPA, and state regulation agencies for conversion into RDF using the Data Conversion component. During the conversion process, data level provenance information for the downloaded and converted data is captured. Next, SWQP loads the converted data into a triple store. When the user accesses the front-end interface of SWQP and issues a request, the request is sent to the back-end reasoning component. The reasoning component then loads the TWC Water Ontology, appropriate regulation ontologies, appropriate water quality data and performs analysis. After the reasoning component completes its analysis, the results are sent to the visualization component for user presentation.

C. *Source Data*

The data sources incorporated into SWQP span several government agencies, including the EPA and USGS, and federal and state regulation agencies.

[9] http://was.tw.rpi.edu/swqp/map.html

[10] http://was.tw.rpi.edu/swqp/trend/epaTrend.html?state=RI&county=3&site=http%3A%2F%2Ftw2.tw.rpi.edu%2Fzhengj3%2Fowl%2Fepa.owl%23facility-110000312135

EPA Data: We obtain permit compliance and enforcement status of facilities regulated by the National Pollutant Discharge Elimination System (NPDES) under the Clean Water Act (CWA) [11] from ICIS-NPDES [12], an EPA system. The compliance and enforcement status of facilities contains measurements of pollutants in the water discharged by the facilities, and also the threshold values for up to five test types for each pollutant. Three test types (C1, C2, C3) use concentration-based limits, while the other two (Q1, Q2) use quantity-based or mass-based limits.

USGS Data: We also fetch the National Water Information System[13] (NWIS) water quality data provided by USGS. The NWIS water quality data provides measurements of substances contained in water samples collected at USGS data-collection stations.

Regulation Data: The water portal makes use of water regulations, which are lists of pollutants and their maximum contaminant level[14] (MCLs). The national level drinking water regulations from EPA, and the state drinking water regulations for California, Massachusetts, New York, and Rhode Island have been encoded and incorporated into SWQP.

III. RESULTS

In this section, we presents how semantic web technologies can serve as useful technologies for solving problems in the domain of water quality investigation from the following aspects: semantic data integration, semantic reasoning, and provenance support.

A. Semantic Data Integration provides an effective and low cost approach for integrating data from various sources.

SWQP integrates data from various sources, including EPA, USGS, and state governments. Our data conversion not only generates ontology-ready RDF data, but also achieves benefits such as aligning terminologies and linking to external resources. For example, we map the property "CharacteristicName" in the USGS dataset and the property "Name" in the EPA dataset to a common property twcwater:hasCharacteristic. What's more, we promote references to characteristic names from string literals to URIs, e.g. "Arsenic" is promoted to "twcwater:Arsenic", which could be linked to external resources like "dbpedia:Arsenic" using owl:sameAs.

The cost of our data conversion is relatively low. We have generated 89.58 million triples for the USGS datasets and 105.99 million triples for the EPA datasets, for CA, MA, NY and RI. For converting the water quality data with csv2rdf4lod, all we need to do is to set up the conversion parameters for each dataset. We converted 139 rules for the MA regulation, 104 for the CA regulation, 100 for the RI regulation, 83 rules for the EPA regulation, and 74 for the NY regulation. The cost of converting the regulations is about 2 person-days for developing the ad hoc converter.

B. Query and reasoning supported by semantic technologies improves responsiveness and simplifies the development of web applications.

The large number of triples generated from the water quality data could cause long response time of the portal. To speed up the reasoning, we use SPARQL queries to narrow down the data and reason over only the relevant data on one selected regulation. In our case, we retrieve and reason only the data for the county corresponding to the input zip code.

Semantic reasoning also eases the complexity of queries a developer needs to write for software applications. For example, to query polluted water sources without reasoning, the web developers need to write complex queries as shown in (1), which compares all measurements from a water source against all limits defined in the regulation. However, with pre-computed results, the developer can simply query polluted water sources and their related information as shown in (2).

```
SELECT * WHERE {
  ?watersource   twcwater:hasMeasurement   ?measurement.
  ?measurement   twcwater:hasValue         ?value;
                 twcwater:hasCharacteristic ?charactericsitc;
                 twcwater:hasUnit          ?unit.                 (1)
  ?regulation    twcwater:hasValue         ?limit;
                 twcwater:hasCharacteristic ?characteristic;
                 twcwater:hasUnit          ?unit.
  ?watersource geo:lat   ?lat;   geo:long ?long.
  FILTER( ?value > limit )
}
SELECT * WHERE {
  ?watersource rdf:type twcwater:pollutedWaterSource.
               geo:lat   ?lat;                                    (2)
               geo:long ?long.
}
```

C. Provenance information encoded using semantic web technology supports transparency and trust.

The primary purpose of SWQP is to discover polluted water sources and polluting facilities in areas a user finds interesting. However, SWQP responses may not be trusted by some users if there is no mechanism that provides the option to examine how the responses are obtained. As pointed out in [12], knowledge provenance, which includes source identification, source authoritativeness, and a supporting graph, can be used to provide explanations. These "explanations" help users understand where responses come from, and what they depend on, thus allowing users to determine for themselves whether they trust the responses they received.

Our portal not only keeps provenance for water quality data, it also keeps provenance for water regulations via the ad hoc converter, which include the URLs of the source, intermediate and converted data, modification time, and source organization. The provenance can be accessed by clicking the question marks in the comparison table [15] of the limits for different pollutants defined in the federal and state water regulations.

[11] http://www.epa.gov/agriculture/lcwa.html

[12] http://www.epa-echo.gov/echo/compliance_report_water_icp.html

[13] http://waterdata.usgs.gov/nwis

[14] http://water.epa.gov/drink/contaminants/

[15] http://tw.rpi.edu/web/project/TWC-SWQP/compare_five_regulation

The user can browse the comparison table to investigate the source of the water regulations and their differences. The user might choose a "what if" scenario, such as to apply a stricter regulation from another state to a local water source. For example, if Rhode Island regulations are applied to water quality data for zip code 02888, 13 polluted water sites are identified. When California regulations are applied, 16 polluted water sites are identified (shown in Fig. 5). Using California criteria on this same region, the indicated number of polluted water sites increases by 23% compared to the number indicated using RI regulation criteria. If we compare the results of using California criteria with using EPA regulations, the number of polluted sites grows by 700%.

SWQP brings together seemingly disparate regulatory and measurement data from multiple sources and, through automated classification and visualization, it can present the data to non-expert users. It provides basic tools to enable users to evaluate exploratory hypotheses. The availability and integration of data are critical to the portal's ability to rapidly disseminate information to the public. With tools such as SWQP, the public could review historical water quality data quickly. Further, citizen scientists could provide their own sample collection and testing data along with its provenance. Although citizen-scientist findings may not be as reliable as experts', they may be timelier and lead authorities to more appropriate testing and validation.

IV. DISCUSSION

Environmental informatics research often benefits from domain knowledge. For example, water quality research requires domain knowledge concerning pollutants, thresholds for pollution, and pollutant test options. Applications that aim to integrate and disseminate water quality data to support analyses related to pollution need to capture and interpret domain knowledge such as sufficient conditions for determining water pollution states and events. Our work is the first we know of that uses a semantic approach to a provenance-aware water quality portal. Other works focus on facilitating water quality management [13, 14] and wastewater treatment [15] via knowledge sharing and reuse. Chen [13] proposed a prototype system that integrates water quality data from multiple sources and retrieves data using semantic relationships among data. Chau [14] presented an ontology-based Knowledge Management system (KMS) that can be integrated into the numerical flow and water quality modeling to provide assistance on the selection of a model and its pertinent parameters. OntoWEDSS [15] is an environmental decision-support system for wastewater management, which augments classic rule-based and case-based reasoning with a domain ontology. SWQP differs from these projects in that it supports provenance based query. For example, users can select to query data only from data sources they trust by selecting them within the Data Source facet. Since SWQP captured provenance information of the data collected, it knows which data came from which sources. This information can then be used to query and visualize only the data from selected sources. Moreover, SWQP is built upon standard semantic technologies (e.g. OWL, SPARQL, Pellet, Virtuoso) and thus can be easily replicated or expanded.

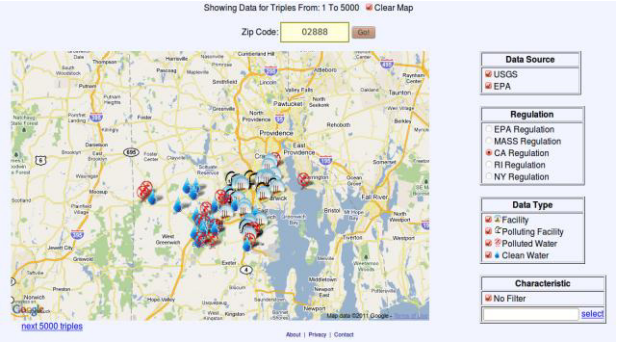


Figure 5. Applying California regulation data to RI water quality Data

SWQP could be expanded in several ways. 1) We can expand SWQP to support all 50 US states. Water quality data can be obtained from EPA and USGS websites. Then, SWQP could identify water pollutions in all the states according to the federal water regulation (or other state regulations we have already encoded such as CA and RI). To convert the remaining state regulations, we could use our existing ad-hoc converters or potentially new converters if the regulations are in different forms. 2) We can quickly add interesting applications to SWQP by integrating data from other sources, e.g. weather and flood forecasts. Flood conditions can exacerbate pollution impacts when pollutant control strategies fail due to floods or when a polluted water source is mingled with a non-polluted water source. If weather conditions suggest anticipated flood regions, SWQP could identify polluting facilities near the flood zone and potentially identify risks and suggest compensating strategies. 3) Another direction is to model the health effects from exposure to the excessive pollutants in water and support reasoning over these effects. Then, SWQP could provide queries customized to health concerns. If the user inputs that he/she is concerned with water pollutants that negatively impact kidneys, SWQP could highlight water sources with high levels of cadmium given the rule that long-term exposure to excessive cadmium may cause kidney damage. 5) The architecture of SWQP can be used for other environment topics. We can build semantic web portals for investigating air quality, soil quality, etc. using the same architecture and workflow used in SWQP. For example, the TWC Clean Air Status and Trends demo [16] has gone through an update to include provenance and could be expanded to include the regulation views. 6) Current SWQP only supports static regulatory levels. However, we captured provenance data about the modification date of the regulations. A web service that regularly checks the updates of the water regulations could recognize when to programmatically download and convert the updates so that the regulations in SWQP are up to date.

As the portal is expanded for greater usage, its credibility becomes more important. To increase the credibility of the portal, we plan to augment its provenance support by building, linking and displaying proof traces that track how the answers are derived from source data. Our PML and Inference Web (IW) provenance infrastructure [16] makes it easy to encode all the data manipulations and use that information for presenting either a complete trace or abstracted trace for user inspection.

[16] http://logd.tw.rpi.edu/demo/clean_air_status_and_trends_-_ozone

We also would like to support provenance granularity options so that users can choose the granularity of the provenance they prefer in certain contexts.

Several e-Science systems have incorporated similar types of provenance support. myGrid [17] proposes the COHSE open hypermedia system, which generates, annotates and links provenance data to build a web of provenance documents, data, services and workflows for biological experiments. The Multi-Scale Chemical Science (CMCS) [18] project developed a general-purpose infrastructure for collaboration across many disciplines. It also contains a provenance subsystem for tracking, viewing and using data provenance. In future work, we intend to leverage the best of these approaches along with domain-specific provenance needs and our IW provenance infrastructure to provide a more water quality-oriented provenance-aware application. We believe that the IW focus on supporting extraction, maintenance and usage of provenance of answers given by web application and services along with the workflow focus of the other systems will provide a nice complement to this work.

Our SWQP evaluation is currently experiential. We demonstrate capabilities that have previously not been possible or not done as efficiently in other architectures. Additionally, we are not aware of a best practice evaluation benchmark for interdisciplinary environmental informatics portals such as this. It is also difficult to evaluate the ontology against existing related ontologies because the driving use case is different and thus the ontologies have significant differences. However, through the design and implementation of SWQP, we have demonstrated the value and potential of applying semantic technologies to facilitate environmental research and community awareness. In the future, we would like to engage both researchers from the hydrology community and interested citizens to evaluate the portal. Feedback from the two user groups can lead to improvements of the portal.

V. CONCLUSION

We presented the TWC Semantic Water Quality Portal, which allows user to discover polluted water sources and polluting facilities. We have illustrated benefits of applying semantic web technologies to water quality research. These benefits include effective integration of heterogeneous data, automatic detection polluted water sources and polluting facilities via semantic reasoning, and increase credibility from utilizing provenance data. We also discussed the extensibility of the portal and the potential of using it for topics beyond water quality. We believe this semantic approach will make it easier to build and maintain environmental informatics portals and empower local communities to track environmental concerns supported by transparent and accessible environmental data.

ACKNOWLEDGMENT

The authors would like to thank the students in the Semantic e-Science 2010 course and Advanced Semantic Technologies 2011 course for their feedback and discussion.

REFERENCES

[1] "Boil Water Advisory Issued for Bristol County Water Authority" Rhode Island Department of Health, September 8, 2009. http://www.ri.gov/press/view/9685

[2] Morgan, T. J. 2009. "Bristol, Warren, Barrington residents told to boil water" Providence Journal, September 8, 2009. http://newsblog.projo.com/2009/09/residents-of-3.html

[3] Raskin, R. G. and Pan, M. J., 2005. Knowledge representation in the semantic web for Earth and environmental terminology (SWEET). Computers & Geosciences 31(9): 1119-1125.

[4] Hobbs, J. R. and Pan, F., 2006. Time Ontology in OWL, W3C Working Draft 27 September 2006. <http://www.w3.org/TR/owl-time/>

[5] Manola, F., Miller, E., McBride, B., 2004. RDF Primer. W3C Recommendation. http://www.w3.org/TR/rdf-syntax/

[6] Lebo, T.; Erickson, J. S.; Ding, L.; Graves, A.; Williams, G. T.; DiFranzo, D.; Li, X.; Michaelis, J.; Zheng, J. G.; Flo- res, J.; Shangguan, Z.; McGuinness, D. L.; and Hendler, J. 2011. Producing and using linked open government data in the twc logd portal (to appear). In Wood, D., ed., Linking Government Data. New York, NY: Springer.

[7] Hitzler, P., Krotzsch, M., Parsia, B., Patel-Schneider, and P., Rudolph, S., 2009. OWL 2 Web Ontology Language Primer. W3C recommendation. <http://www.w3.org/TR/owl2-primer/>

[8] Prud'hommeaux, E., and Seaborne, A., 2008. SPARQL Query Language for RDF. W3C Recommendation 15 January 2008. http://www.w3.org/TR/rdf-sparql-query/.

[9] Sirin, E., Parsia, B., Cuenca-Grau, B., Kalyanpur, A., and Katz, Y. 2007. Pellet: A practical OWL-DL reasoner. Journal of Web Semantics 5(2): 51-53. doi:10.1016/j.websem.2007.03.004

[10] Carroll, J.J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A., and Wilkinson, K. 2004. Jena: Implementing the semantic web recommendations. Proceedings of the 13th International World Wide Web Conference, pp. 74-83. doi: 10.1145/1013367.1013381

[11] McGuinness, D.L., Ding, L., Pinheiro da Silva, P., and Chang, C. 2007. PML 2: A Modular Explanation Interlingua. Workshop on Explanation-aware Computing, July 22-23, 2007.

[12] Pinheiro da Silva, P., McGuinness, D.L., and McCool, R. 2003. Knowledge Provenance Infrastructure. IEEE Data Engineering Bulletin, vol. 26.

[13] Chen, Zhiyuan, Gangopadhyay, Arrya, Holden, Stephen H., Karabatis, George, McGuire, Michael P., 2007. Semantic integration of government data for water quality management. Government Information Quarterly 24(4): 716–735. doi: 10.1016/j.giq.2007.04.004.

[14] Chau, K.W., 2007. An Ontology-based knowledge manage ment system for flow and water quality modeling. Advances in Engineering Software 38(3): 172-181.

[15] Ceccaroni, L., Cortes, U. and Sanchez-Marre, M., 2004.OntoWEDSS: augmenting environmental decision-support systems with ontologies, Environmental Modelling & Software 19(9): 785-797.

[16] McGuinness, D. L., and Pinheiro da Silva, P., 2004. Explaining answers from the semantic web: The inference web approach. Journal of Web Semantics 1(4):397-413.

[17] Zhao, J., Goble, C. A., Stevens, R. and Bechhofer S., 2004. Semantically linking and browsing provenance logs for e-science. Semantics of a Networked World 3226: 158-176. doi: 10.1007/978-3-540-30145-5_10

[18] Myers, J., Pancerella, C., Lansing, C., Schuchardt, K., and Didier, B., 2003. Multi-scale science: Supporting emerging practice with semantically derived provenance. ISWC workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data.

Birds of a Feather Sessions

1. Geospatial Data Management for Ecological Research Organizations

Theresa Valentine[1], Adam Skibbe[2], Jamie Hollingsworth[3]
[1]US Forest Service Research Branch, [2]Konza Prairie LTER, Kansas State University, [3]Bonanza Creek LTER, University of Alaska Fairbanks,
Theresa.valentine@oregonstate.edu, askibbe@ksu.edu, jhollingsworth@alaska.edu

The management of geospatial data has traditionally been conducted within a separate Geographic Information System (GIS). Improvements in the interoperability of these systems with traditional data management systems have resulted in improved integration with place based data records.

Improvements in user interfaces, such as Google Earth and other internet mapping applications, and the availability of low cost GPS receivers have increased the public's awareness and use of place based data. The integration and interoperability of these data are becoming critical for the synthesis of data within and between different ecological sites and programs.

This session will focus on the challenges and opportunities that information managers encounter with the increase in demand and in volume of geospatial data and integrating this data with research data collected as part of field studies. Areas of interest may include providing access and analysis tools for large LiDAR datasets, documenting geospatial data within FGDC and EML metadata content standards, developing and managing Citizen Scientist data, and strategies to obtain study site locations and use these locations to provide geographic searches using mapping tools.

The session welcomes anyone interested in managing Geospatial data, using open source or proprietary software options. Outcomes will include a summary analysis of best approaches to different data tools as well as an outline document of possible tools for Ecological Information Managers. Attendees will be encouraged to share demos, experiences and projects that might be of interest to the group.

2. Internet Mapping: What are the options?

Jamie Hollingsworth
Bonanza Creek LTER Site,
Jhollingsworth@alaska.edu

Currently, there is an increase in demand for serving and displaying various types of spatially referenced data. As the industry moves further towards cloud based storage and computing and Internet-based applications, the number of options available for data managers to communicate information has grown quickly. In this session, we will discuss some of the options available for serving and displaying spatially referenced information. We will also talk about challenges in displaying these data and associated metadata on the internet. Specific topics may include: how to pick an appropriate software package, determining and analyzing various data source types (e.g. dynamic versus static), and the extensive variety of tools a user may have access to.

In this session, we welcome input and discussion from those interested in internet mapping, regardless of experience. We encourage participation from data managers currently looking for ideas on how to create internet mapping products as well as more advanced programmers willing to share experiences and insight.

The outcome of this session will be four-fold. Firstly, we expect to discuss and record shared experiences. Secondly, we will present demos of existing internet mapping tools, and we will explore existing services. We will educate interested data managers regarding what internet mapping options currently exist. Finally, we will discuss future applications of internet mapping within the context of ecological data management.

Session Length: This session will be two hours in length. 45 minutes will be spent discussing what internet mapping options are currently available, one hour be used to educate and demonstration how internet mapping applications work and what it takes to create one, and 15 minutes will be spent looking towards the future of internet mapping.

3. Using Web Tools and Methods to Support Earth Science Collaborations

Erin Robinson
Foundation for Earth Science
erinrobinson@esipfed.org

Many Earth science projects have participants that span multiple timezones, organizations and domains. Sometimes members of the group have never even met face to face. The requirement to be co-located in order to collaborate is no longer the norm since there are now so many alternative methods of virtual communication and coordination using web tools and methods. There are many tools (Drupal, Mediawiki, Google +, Twitter, Facebook) that support communication, coordination and collaboration around a topic. The good thing about all of these tools is that they are flexible and customizable, but this also poses a challenge of how to set-up the tools to best support your group. Often these collaborations are supported by an ad-hoc member of the group, who is working within the group, but also is supporting the collaboration of the group. This person often will have created methods to supporting the group such as sending out the email reminders, hosting the telecons and updating the web pages. This at times can be a frustrating job because only a small fraction of the group participates at any given time. This Birds of a Feather session is intended to bring together these ad-hoc community manager practitioners to compare what is working to support virtual collaboration and what are the challenges. Hopefully, the outcome of this session will be a web-based forum to improve the efficiency of these Earth science community managers.

4. Automating Data Processing and Quality Control using Workflow Software: Converting Sensor Data to Usable Environmental Information

Wade Sheldon[1], John Porter[2]
[1]Georgia Coastal Ecosystems LTER, [2]Virginia Coast Reserve LTER
wsheldon@lternet.edu, jporter@lternet.edu

Advances in sensor technology, computer hardware and wireless networking make it possible to measure a wide range of environmental variables simultaneously across multiple temporal and spatial scales, and acquire the data in real time. These advances present exciting new research opportunities, but they have also led to dramatic increases in the volume of data that can be acquired by environmental research projects. Processing, documenting and quality controlling high-volume data sets is a major challenge for many environmental information managers due to poor scalability of traditional interactive approaches. Strategies for automating these operations are clearly needed.

Open source scientific workflow applications (e.g. Kepler, https://kepler-project.org/), streaming data engines (e.g. DataTurbine, http://www.dataturbine.org/) and metadata-driven data processing software (e.g. GCE Data Toolbox, https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox) hold great promise for automating data processing and quality control to turn raw sensor data into usable environmental information. However, many barriers exist (both real and perceived) to adopting these tools in the EIM community. We propose to conduct a 2 hour "Birds of a Feather" session at the 2011 EIM Conference to explore this issue through a combination of 2-3 short software demonstrations and a 1 hour round-table discussion.

Demonstrations will illustrate real-world use of workflow software to process and quality control environmental sensor data, with an emphasis on computer requirements, first steps, and resources for getting started. The round-table discussion will focus on: 1) Overcoming barriers to adoption of workflow software, 2) Strategies for finding pre-built workflows, collaborators and training, 3) Striking the right balance between general (sharable but time-consuming) and specific (proprietary but easier) approaches, and 4) What we still can't do with workflow software (i.e. development needs).

We believe that this session will educate EIM participants about effective strategies for automating data processing using workflow approaches, and will foster collaboration on workflow development and sharing within the EIM community. Another potential outcome will be development of a broader workshop or training proposal to NSF or the LTER network.

5. Functional Requirements for the EML Dataset Congrence Checker

Margaret O'Brien[1], Mark Servilla[2]
[1]University of California Santa Barbara, Santa Barbara Coastal LTER, [2]University of New Mexico, LTER Network Office,
mob@msi.ucsb.edu, mservilla@lternet.edu

Ecological Metadata Language (EML) is a widely-used specification for metadata describing environmental data resources (e.g., data tables). Workflows and other automated processing tools must be able to use these metadata documents to access and process those resources, but experience indicates that a significant fraction of available EML data entities are not of sufficient quality for this use. Currently few tools exist for assessing data-metadata agreement (congruence), and the LTER Network has outlined this need (O'Brien et al, 2009). The LTER Network has begun work on software tools for assessing and reporting on the usability of EML datasets for automated loading and processing (the EML Congruence Checker) as part of the suite of Network Information System web services. The software extends the EML Data Manager Library, Java code which parses EML metadata documents and handles data entities using relational database constructs. The Congruence Checker project finished its first development cycle during mid-2011. An initial set of checker requirements was developed and categorized according to system, type and return-status, and an initial report format proposed. The LTER EML metrics and congruence group would like to introduce this work to a broader audience and solicit feedback. Format and target participants: The session is planned to include a short (15 minute) demonstration of the ECC web services and reports, a summary of the project status, and the developing rationale behind the classification of the list of checks. The bulk of the session is to be a discussion, e.g., of current planned requirements, reporting scheme, priorities and possible integration with the Data Manager Library code. EIMC participants comprise a broad spectrum of developers, information managers and specialists, many of whom use EML for data management. The target participants for this session include members of the EML development committee, information managers and others creating or reading EML datasets, developers using the EML Data Manager Library, and users of other metadata specifications considering similar tools.

Plenary Discussion

Community Standards and Practices Development

Organizer Margaret O'Brien

Synopsis: Our strengths as a community of practice are both enhanced and complicated by the diversity of data, and to accomplish broad data synthesis goals it would benefit practitioners to also synthesize the knowledge systems which model and manage those data. Development of shared tools can frequently be simplified if data management and modeling practices converge. The concept of common practices covers many areas of data management. For example, the flexibility of the EML specification has lead to a variety of construction patterns, and common patterns would simplify development of EML‐related applications. Practices for sharing knowledge definitions and logic could be developed (e.g., vocabularies and ontologies), so that knowledge models from many domains can be combined or reused rather than reinvented. Some experience has been gained by developing web services to deliver content in common specifications, and these experiences may be leveraged in other areas.

Anyone considering a new (or upgraded) information management system should be aware of existing data models before designing new custom models, and yet no adequate exchange mechanism currently exists. This session would be used to discuss and promote activities which lead directly to sharing and reusing data models and other software tools. This discussion follows previous work at various venues, and among data practitioners already using common models and practices. Follow‐up activities may include a white paper outlining possible partnerships between groups of practitioners, or working groups agreeing to meet at other upcoming meetings.

Posters

1. The Open Source DataTurbine (OSDT) Android Sensor Pod: Embedded Cyberinfrastructure for Smart Buoy Controllers and Experiments in Ocean Acidification and Limnology

Peter Arzberger[1], Tony Fountain[1], Sameer Tilak[1], Peter Shin[1], Gesuri Ramirez[1], Tim Kratz[2], Corinna Gries[2], Sally Holbrook[3], Russell Schmitt[3], Andrew Brooks[3], Keith Seydel[3], Robert Carpenter[4], Jennifer Smith[5], Todd Martz[5], Matthew Miller[6], John Wilson[6]

[1]University of California, San Diego, [2]University of Wisconsin, Madison, [3]University of California, Santa Barbara, [4]California State University, Northridge, [5]Scripps Institution of Oceanography, UCSD, [6]Erigo Technologies

parzberg@sdsc.edu, tfountain@ucsd.edu, sameer@sdsc.edu, pshinn@ucsd.edu, gesuri@gmail.com, tkkratz@wisc.edu, cgries@wisc.edu, holbrook@lifesci.ucsb.edu, schmitt@lifesci.ucsb.edu, brooks@msi.ucsb.edu, seydel@msi.ucsb.edu, robert.carpenter@csun.edu, smithj@ucsd.edu, trmartz@ucsd.edu, matt.miller@erigo.com, john.wilson@erigo.com

Increasingly, environmental observing systems have become important tools used to understand key environmental processes. These systems require sophisticated cyberinfrastructure (CI) that must be easy to deploy and maintain. The OSDT Android Sensor Pod project is a collaboration between computer scientists, ecologists, and marine scientists to develop cyberinfrastructure (CI) for applications in marine biology and limnology. The project is motivated by requirements from the NSF Long Term Ecological Research Network (LTER), the Coral Reef Environmental Observatory Network (CREON), and the Global Lake Ecological Observing Network (GLEON). The goal is to enable new types of science experiments for real-time, fine-scale monitoring in coral reefs and lakes by making system deployment and operations more efficient. The core of this new CI is an embedded controller for buoy management. Funded by the Gordon and Betty Moore Foundation, this project involves a combination of software and hardware developments together with field deployments at the NTL LTER site (Great Lakes), the MCR LTER site (Moorea, French Polynesia), and Palmyra Atoll (Central Pacific). As part of the project, we ported the OSDT middleware to the Android platform and developed new software for configuring and managing real-time embedded applications. By employing standard Android operating system, the developed software is readily available on a broad range of devices including smartphones, tablets, and netbooks. In essence, choice of Android platform allows us to leverage the tremendous engineering investment made in producing what has become commodity embedded systems. The OSDT-Android controller communicates with sensors through the Sea-Bird Inductive Modem interface and manages sensor interfaces, data acquisition, on-board processing, and data transmission over multiple types of radios, including Iridium satellite, cellular, Bluetooth, and long-distance wireless. When combined with a Droid cell phone or tablet, the controller becomes a robust sensor pod that can be configured to serve as a cluster head or gateway node in complex sensor-based systems. Developed in Java, the OSDT sensor pod can manage a local constellation of sensors and communicates with other OSDT-enabled platforms. It can be readily updated to incorporate new software modules, and dynamically reconfigured to schedule these modules to control sensor operations and communications. It is designed to replace “dumb” data loggers and buoy controllers, support on-platform event detection and real-time control and includes necessary software for scheduling sensor operations and communications. Initial lab tests have been successful. Field deployments are scheduled for Fall 2011, Winter 2012, and Summer 2012. A variety of sensor types, including pH, pCO2, dissolved oxygen, temperature, and pressure, will be utilized during field deployments to investigate several important ecological questions including: (1) How sensitive are ocean systems to pH changes? and (2) What is the variability of lake metabolic parameters such as gross primary productivity and respiration?

Keywords: Open Source DataTurbine, Android Sensor Pod, Ocean Acidification, Coral Reefs, Limnology, Environmental Observing Systems

2. From Fisheries Studies to Biodiversity Data Sharing

Julien Barde
Exploited Marine Ecosystems, EME-212 research unit, 65 avenue Jean Monnet, 34203 Sète Cedex 5, France.
julien.barde@ird.fr

In addition to target species, Ecosystemic approach to Fisheries (EAF) aims to take into account other ecosystem components related to them from an ecological point of view (their preys, seamounts..). Additional information resources are needed to make this new approach effective. Data sharing and then interoperability is a key point required at the beginning of the process. This poster presents our ongoing work on interoperability to make tuna fisheries data available for different kinds of users. The first goal is to facilitate datasets discovery and then, by implementing different data formats and related access protocols, make them understandable and usable by different communities. We choosed standards sets relevant for geospatial data (OGC), biodiversity data (TDWG), statistical data (SDMX), fisheries data (COST). Moreover, the ability to describe, manage and serve our data by mapping their content with these standards requires the management of semantic issues. A new system has been set up, driven by an ontology (using RDF, OWL and SPARQL languages related to semantic Web activity of W3C), which enables to summarize and manage both knowledge and data on ecosystems related to tropical tuna. This knowledge base is made accessible through a Web portal (www.ecoscope.org) and can be used to visualize relationships between components of these ecosystems (like foodwebs). The underlying ontology can be used as a semantic agent to convert terms, referential codes according to the standards and user's profile.

Keywords: Ecosystem Approach to Fisheries, ontologies, metadata, biodiversity, interoperability

3. Cyberinfrastructure for the Tropical Ecology Assessment and Monitoring (TEAM) Network

Chaitanya Baru[1], Eric Fegraus[2], Jorge Ahumada[2], Sandeep Chandra[1], Kate Kaya[1], Kai Lin[1], Choonhan Youn[1]
[1]San Diego Supercomputer Center, [2]Conservation International
baru@sdsc.edu, efegraus@conservation.org, jahumada@conservation.org, chandras@sdsc.edu, kate@sdsc.edu, klin@sdsc.edu, cyoun@sdsc.edu

The Tropical Ecology Assessment and Monitoring (TEAM) Network (www.teamnetwork.org), organized through the partnership of Conservation International, Missouri Botanical Garden, Smithsonian Institution and the Wildlife Conservation Society, is a multi-disciplinary network for monitoring long-term trends in biodiversity and ecosystem services through a network of tropical field sites using scientifically accepted standardized monitoring protocols. The TEAM framework is designed to address a set of grand challenge questions that are fundamental to understanding the dynamics of biodiversity, ecosystem services, and human well-being, as they interact from local to global scales in the context of multiple changing drivers, e.g., climate change and land cover change. TEAM sites are located in the three major continental blocks of humid tropical forests and within those, span a range of latitudes and current and future projected environmental (e.g., climate) and anthropogenic (land use and climate change) gradients. The network provides data on climate, tree and liana species diversity and bird and mammal species diversity.

To support TEAM objectives, a comprehensive cyberinfrastructure is needed with related services for collection, management and dissemination of large amounts of data; timely analysis of the data; and, integration of observational data with other related datasets. The cyberinfrastructure must provide the overall capability to scale-up earth observation information from the currently prevalent mode of small, individual investigator-based data collection and experimentation to larger, multi-institution, multi-disciplinary networks operating at national, regional and global spatial scales.

Data collection and acquisition activities are based on standardized protocols that define the field methodology, spatial and temporal resolutions, and minimum data QA/QC standards, regardless of the means by which data are collected. The cyberinfrastructure developed in TEAM supports fully automated data collection (Climate Sensors), semi-automated data collection (Terrestrial Vertebrate/Camera Traps) and fully manual data collection (Vegetation). To support the varying nature of data collection in the network, cyberinfrastructure tools have been developed to assist field personnel at TEAM sites. The climate data management tool tracks all climate sensor data, including maintenance information such as calibrations and regular checkups. The vegetation data management tool allows users to enter data with the correct taxonomy authorities and validates manually collected data with predefined rules and the previous year's census. A desktop application, DeskTEAM, manages camera trap data by loading photos from TEAM camera traps into a local repository,

identifying animals in the photos using correct genus and species names, and exporting photos to a central database.

Keywords: Tropical Ecology, Cyberinfrastructure, Data Management, Climate, Vegetation, Camera Trap

4. From Rolling Deck to Repository (R2R): Creating a National Pipeline for Underway Shipboard Data

Suzanne Carbotte[1], Stephen Miller[2], Andrew Maffei[3], Shawn Smith[4], Robert Arko[1], Vicki Ferrini[1], Karen Stocks[5], Cynthia Chandler[3], Mark Bourassa[4], Dru Clark[2], Suzanne O'hara[1], Aaron Sweeney[2], John Morton[1]

[1]Lamont-Doherty Earth Observatory, [2]Scripps Institution of Oceanography, UCSD, [3]Woods Hole Oceanographic Institution, [4]Florida State University, [5]San Diego Supercomputer Center

carbotte@ldeo.columbia.edu, spmiller@ucsd.edu, amaffei@whoi.edu, smith@coaps.fsu.edu, arko@ldeo.columbia.edu, ferrini@ldeo.columbia.edu, kstocks@ucsd.edu, cchandler@whoi.edu, bourassa@coaps.fsu.edu, pdclark@ucsd.edu, sohara@ldeo.columbia.edu, asweeney@ucsd.edu, jmorton@ldeo.columbia.edu

Launched in 2009, R2R is a systematic effort to capture, catalog and archive US underway shipboard data. Each vessel in the US academic fleet is equipped with a multidisciplinary suite of sensors that are available for continuous operation during each expedition. The "underway" geophysical, water column, and meteorological datasets obtained from these sensors describe basic environmental conditions for the oceans and are of high value for building global syntheses, climatologies, satellite validation and historical time series of ocean properties. The R2R Portal (www.rvdata.us) will be the central gateway through which underway data are routinely cataloged and securely transmitted to the appropriate national data center, ensuring long-term access and relieving chief scientists of their individual obligations under NSF policy to submit underway data.

Protocols are being developed for quality assessing high priority underway data types, to provide feedback to shipboard instrument operators and inform end users. Standard metadata will be supplied with each dataset, including provenance and quality information. Standard products, such as quality-controlled navigation, are being created. As part of this work, R2R is collaborating with NOAA to create an XML-based, ISO 19115-compliant cruise metadata template. This describes the basic elements of a seagoing expedition: cruise identifier, vessel name, operating institution, dates/ports, navigation track, survey targets, science party, funding sources, scientific instruments, daughter platforms, and data sets. Controlled vocabulary terms are directly embedded as Uniform Resource Identifier (URI) references. We envision a hierarchical framework where a single "cruise-level" record is linked to multiple "dataset-level" records that may be published independently.

One of the subprojects within R2R is the development of a shipboard scientific event logging system that incorporates best practice guidelines, controlled vocabularies, a cruise metadata schema, and a scientific event log. The ELOG-based cruise event logging system, currently being tested, enables researchers to record digitally all scientific events and assign a unique event identifier to each entry, to assist in the ingestion of these data into oceanographic data repositories and subsequent reuse of the datasets.

As of July 2011, data from 2,130 cruises on 26 vessels had been submitted, totaling 7,481,290 files (>9 TB).

Rolling Deck to Repository is a collaboration between Lamont-Doherty Earth Observatory (lead institution), Scripps Institution of Oceanography, San Diego Supercomputer Center, Woods Hole Oceanographic Institution, and Florida State University; and works with the vessel operating institutions, UNOLS Office, NOAA National Data Centers, and disciplinary data assembly centers (DACs).

Keywords: Ocean informatics, Oceanography, Metadata, Quality control

5. Metadata management in NASA's Earth Observing System (EOS) ClearingHOuse (ECHO)

Matthew Cechini[1], Andrew Mitchell[2]

[1]Raytheon - NASA ESDIS, [2]NASA – ESDIS

Matthew.F.Cechini@nasa.gov, Andrew.E.Mitchell@nasa.gov

Metadata is an important entity in the process of cataloging, discovering, and describing earth science data. As science research and the gathered data increases in complexity, so does the complexity and importance of descriptive metadata. To meet these growing needs, the metadata models required utilize richer and more mature metadata attributes. Categorizing, standardizing, and promulgating these metadata models to a politically, geographically, and scientifically diverse community is a difficult process.

Whether a Earth Science Data System (ESDS) or an independent Principle Investigator, each finds itself or themselves responsible for navigating the difficult realm of metadata management. When dealing with metadata, there are at least 4 core activities or responsibilities that must be addressed:

- Generation – The process whereby metadata is generated according to a specific format or standard with the appropriate processing lineage.
- Data Discovery – The utilization of metadata to correlate and discover data based on information provided within a metadata record.
- Retrieval – The process of accessing metadata in its native or a translated format for viewing or further utilization.
- Preservation – The short and long-term archival of metadata to ensure its accessibility for future needs.

NASA's Earth Observing System Data and Information System (EOSDIS) is a complex Earth Science Data System comprised of 12 data centers, each focusing on a separate scientific domain of research. The EOSDIS addresses each of the identified core metadata activities. Each Data Center has primary responsibility for metadata generation and preservation, acting as the curators for their data holdings. An integral component of metadata management within the EOSDIS is NASA's Earth Observing System (EOS) ClearingHOuse (ECHO). ECHO is the core metadata repository for the EOSDIS data centers providing a centralized mechanism for metadata and data discovery and retrieval.

NASA's EOSDIS has taken special interest in investigating the adoption of the International Standards Organization's (ISO) 19115/19/39 metadata standard. Moving to adoption of a new standard requires significant modifications to internal metadata workflows in each of the 4 areas of ingest activity.

ECHO has undertaken an internal restricting to meet the changing needs of scientists, the consistent advancement in technology, and the advent of new standards such as ISO 19115. These improvements were based on the following tenets for data discovery and retrieval:

- There exists a set of 'core' metadata fields recommended for data discovery.
- There exists a set of users who will require the entire metadata record for advanced analysis.
- There exists a set of users who will require a 'core' set metadata fields for discovery only.
- There will never be a cessation of new formats or a total retirement of all old formats.
- Users should be presented metadata in a consistent format of their choosing.

In order to address the previously listed items, ECHO's new metadata processing paradigm utilizes the following approach:

- Identify a cross-format set of 'core' metadata fields necessary for discovery.
- Implement format-specific indexers to extract the 'core' metadata fields into an optimized query capability.
- Archive the original metadata in its entirety for presentation to users requiring the full record.
- Provide on-demand translation of 'core' metadata to any supported result format.

With this identified approach, the Earth Scientist is provided with a consistent data representation as they interact with a variety of datasets that utilize multiple metadata formats. They are then able to focus their efforts on the more critical research activities that they are undertaking.

Keywords: ECHO, Metadata, NASA, ISO 19115, Data Discovery

6. NASA's EOSDIS Coherent Web Platform Development

Matthew Cechini[1], Kevin Murphy[2], Greg Baerg[1]
[1]Raytheon - NASA ESDIS, [2]NASA ESDIS
Matthew.F.Cechini@nasa.gov, kevin.j.murphy@nasa.gov, Gregory.A.Baerg@jpl.nasa.gov

The Earth Observation System Data and Information System (EOSDIS) provides valuable data and services to a global Earth Sciences community. The twelve EOSDIS data centers allow for focused attention on the various unique science disciplines. While each data center has an independent identify, each is also a part of the broader EOSDIS community. In addition to data centers, the EOSDIS community includes a broad array of metadata clearinghouses, data services, user working groups, standards organizations, collected metrics, and system interfaces. Individually, each system component serves an important function and role, however, it is the aggregation of components that brings enhanced value to the Earth Science community.

The ESDIS Project, which has management responsibility for EOSDIS, is undertaking an effort to create a consistent presence for EOSDIS, dubbed the "Coherent Web" Project. The first phase (Phase I) Coherent Web activities, scheduled for completion in late 2011, has the following goals in mind:

- Create a consolidated website pulling together existing content into a single location;

- Create a top-hat navigation bar for inclusion in all EOSDIS data center websites, improving the community association;
- Create a programmatic structure and workflows for managing and approving content;
- Create a methodology and platform where additional content and services can be incorporated or hosted for end-user access; and
- Identify Phase II activities to provide continual improvements to EOSDIS information and data discovery.

The overall approach, goals, and achievements of the Phase I Coherent Web activities will be presented in poster form.

Keywords: Drupal, Platform, Coherent Web, EOSDIS, NASA

7. NASA Reverb: Metadata-Driven Earth Science Data & Service Discovery

Matthew Cechini[1], Andrew Mitchell[2]
[1]Raytheon - NASA ESDIS, [2]NASA – ESDIS
Matthew.F.Cechini@nasa.gov, Andrew.E.Mitchell@nasa.gov

NASA's Earth Observing System Data and Information System (EOSDIS) is a core capability in NASA's Earth Science Data Systems Program. The EOSDIS contains 12 data centers each responsible for stewardship over separate scientific domains. A core function of the EOSDIS is to facilitate the discovery, access, and interpretation of data that is collected by the EOSDIS. NASA's EOS ClearingHOuse (ECHO) is a metadata catalog for the EOSDIS data centers, providing a centralized catalog of data products and registry of related data services.

Earth scientists can access EOSDIS data and services by using general or community-tailored clients that access ECHO's data and service holdings. WIST, the Warehouse Inventory Search Tool, has been the primary web-based client for discovering and ordering cross-discipline data from all of ECHO's metadata holdings for many years and has served the Earth Science community well. Working closely with this community, the ECHO team identified a need to develop the next generation EOS data and service discovery tool.

The ECHO Team based their client development efforts on the following principles:

- Metadata Driven User Interface – Users should be presented with data and service discovery capabilities based on dynamic processing of metadata describing the targeted data.
- Integrated Data & Service Discovery – Users should be able to discovery data and associated data services that facilitate their research objectives.
- Leverage Common Standards – Users should be able to discover and invoke services that utilize common interface standards.

After a yearlong design, development, and testing process, the ECHO team successfully released "Reverb – The Next Generation Earth Science Discovery Tool." Reverb was developed in a fast-paced agile development process requiring constant interaction between the developers, product owners, customers, and end-users. Reverb provides a success story of close community involvement to produce an enhanced earth science discovery platform.

Metadata plays a vital role facilitating data and service discovery and access. As data providers enhance their metadata, a user's search experience may also be enriched, as they are able to discover items of interest using more advanced search capabilities. Reverb's reliance on metadata provides a dynamic experience to users based on identified search facet values extracted from science metadata. Utilizing cross-dataset correlation and search based on provided metadata values, users can discover additional dataset that they may not previously have been aware of.

Data discovery and access is not limited to simply the retrieval of data granules, but is growing into the more complex discovery of data services. These services include, but are not limited to, services facilitating additional data discovery, subsetting, reformatting, and re-projecting. The discovery and invocation of these data services is made significantly simpler through the use of consistent and interoperable standards. Sample standards include the OGC and OPEnDAP protocols. By utilizing an adopted standard, developing standard-specific adapters can be utilized to communicate with multiple services implementing a specific protocol.

Through Reverb, users may discover services associated with their data of interest. When services utilize supported standards and/or protocols, Reverb can facilitate the invocation of both synchronous and asynchronous data processing services. This greatly enhances a users ability to discover data of interest and accomplish their research goals.

Extrapolating on the current movement towards interoperable standards and an increase in services, the ultimate goal is to provide a ubiquitous experience for users when discovering data. Services will become a natural part of data discovery, reducing users needs to be aware of the service that is facilitating their data access. The Reverb discovery tool provides a platform to shift the earth science data discovery paradigm.

Keywords: Service, Standards, NASA, ESDIS, EOSDIS, Reverb, Client, Data Discovery

8. An Educational Experiment Using an Ecology Data Archive -- Making the Most of Metadata, Reproducing Results, and Training Students for Synthesis Science

Judith Bayard Cushing, Kathleen Saul
The Evergreen State College
judyc@evergreen.edu, saulk@evergreen.edu

In February, 2011, Victoria C. Stodden (Columbia University) organized a symposium at the American Association for the Advancement of Science Conference that addressed Reproducibility and Interdisciplinary Knowledge Transfer. While that symposium dealt primarily with computational results, her contention that the inherent difficulty in verifying published results of computational research might be "leading to a credibility crisis affecting many scientific fields" sparked the authors to consider the questions:

Could ecology data archives be used to verify research results? If so, what methods would be most effective for accomplishing this?

This poster will report on an educational experiment conducted in our Spring 2011 Master's level course in quantitative methods, where we asked our 27 students to work in teams of 2-3 to conduct an analysis on one or more data sets from the H.J. Andrews Experimental Forest, using both metadata and published results to guide that analysis. Students were free to conduct "new" research, or attempt to reproduce some of the reported results. Other questions that motivated this experiment were:

1. Quantitative methods and statistics texts typically focus on concepts. While this is right-minded, textbook data have usually been greatly shortened, simplified, and sanitized. As a result students rarely leave the course understanding how to use, manage, validate, or document large data sets – until they collect the data themselves which often leads to painful experiences and losses in data, time, and money. How might we provide realistic experiences with large data sets prior to students' own research?

2. How might we train students to integrate data and conduct synthesis science?

3. How might future researchers get into the habit of effectively using ecology data archives, a skill that will become increasingly important as the cost of collecting data rises, and the ecological observatories come on board.

4. How useful are existing metadata in helping scientists use the associated data? Which are the most useful metadata? Where are the stumbling blocks?

Keywords: Training for Synthesis Science, Reproducing Scientific Results, Metadata Use

9. A Generative, Multisensor Model for Quality Control in Ecological Data

Ethan Dereszynski, Thomas Dietterich
Oregon State University
ewdere04@yahoo.com, tgd@eecs.orst.edu

Contemporary environmental science is increasingly reliant upon networks of distributed automated sensors in remote locations. Decreased cost and improved portability of these sensors have allowed researchers to monitor landscapes at very fine spatial and temporal granularities. An instrumented research site may generate dozens to hundreds of near-continuous data streams of environmental measurements. However, in-situ sensors are often subject to harsh conditions that can lead to malfunctions in individual sensors and failures in network communications. Quality control (QC) is essential to identify incorrect measurements before these data can be assimilated in models and analyses. However, the abundance of data makes manual inspection by domain experts impractical and delays the release of data.

In this poster, we describe a generative modeling approach to automated QC. A probabilistic approach is provided that allows us to maintain a distribution over the functioning state of a sensor and the true value of the monitored phenomena. This framework facilitates real-time QC wherein we simultaneously diagnose the working state of the sensor and infer a distribution over its current reading. We explore machine learning techniques for learning the joint relationship among different types of sensors at a monitoring site. Our model is evaluated using three meteorological stations deployed in the H.J. Andrews Forest, a Long-term Ecological Research (LTER) site in western Oregon. We compare our results to existing single and multiple-sensor QC models.

Keywords: Quality Control, Sensor Networks, Bayesian Modeling, Machine Learning

10.The BIOTA/FAPESP Program: The Virtual Institute of Biodiversity

Debora Drucker[1], Tiago Estrada[2], Carlos Joly[2] and José Salim[2]
[1]EMBRAPA, [2]UNICAMP
debora@cnpm.embrapa.br, tiagode@unicamp.br, cjoly@unicamp.br, jose.asalim@gmail.com

Since the Convention on Biological Diversity (CBD) in 1992, biodiversity conservation (the protection of species, ecosystems, and ecological processes) and restoration (recovery of degraded ecosystems) have been high priorities for many countries. CDB highlights the importance of making information on biodiversity knowledge available to researchers, policy-makers and the general public. This prompted scientists in 1999 to found the Virtual Institute of Biodiversity, BIOTA-FAPESP. FAPESP, the State of São Paulo Research Foundation, is a nonpolitical, taxpayer-funded foundation, one of the main funding agencies for scientific and technological research in Brazil, and a supporter of this program. During its first 10 years, the program supported 94 major research projects, described more than 1800 new species, acquired and archived information on over 12,000 species, and made data from 35 major biological collections available online, a first for Brazilian biological collections. One of the challenges for the next 10 years of BIOTA-FAPESP is to improve its information system in order to also accommodate ecological data, linked to the taxonomic data. Researchers from BIOTA/FAPESP Functional Gradient Project are testing tools to advance on Biodiversity Information Management in order to enhance the long-term value of existing data by making it available for further research. Advances include a generic and spatial enabled database system to accommodate field survey data. All information is documented in Ecological Metadata Language (EML). The data packages (datasets accompanied by metadata documentation) are stored in a Metacat instance in the Institute of Biology, UNICAMP. Our effort is a contribution to advance on biodiversity information integration and to foster synthetic studies.

Keywords: Biodiversity, Data sharing, Ecology

11.Near-Real Time Anomaly Detection for Eddy Covariance Data: A Case Study

Irbis Gallegos
The University of Texas at El Paso
irbisg@miners.utep.edu

Eddy covariance (EC) methods are used to measure exchange of carbon dioxide (CO2), water vapor and energy between land and atmosphere. The amount of eddy covariance data being collected by sensor towers at long-term ecological research sites is increasing and the ability to evaluate the accuracy of the data at near-real time and to check that the instrumentation is operating correctly become critical in order to not lose valuable time and information. This poster presents an approach to specify and verify data properties that detect anomalies in EC sensor data. In this context, an anomaly is a deviation from an expected sensor data value or data behavior.

For this work, scientists used the Data Property Specification (DaProS) prototype tool to specify, refine, and validate data properties of interest based on existing expert-knowledge, algorithms, and protocols used by EC scientific communities. DaProS is a scientist-centered tool that assists the user in specifying a data property through a series of guiding questions and selections. The tool yields the appropriate specification as well as a disciplined natural language representation of the specification for validation purposes.

Datasets of EC data were extracted from the data repository of a newly-placed EC sensor tower located at the LTER Jornada Basin Experimental Range. The selected datasets were of interest because the measurements were taken during the EC tower installation period and during periods of time from the summer and winter seasons when unusual weather events, such as unexpected snow and rain, took place. The continuously-collected EC tower datasets were manually split into 1-hour interval files to simulate near-real time data collection, and were used as input to the Sensor Data Verification (SDVe) prototype tool. SDVe uses DaProS-generated data properties to detect anomalies in scientific sensor data at near real time, or as soon as the data is available from the data logger in the field.

The approach in this work allowed scientists to identify environmental variability, instrument malfunctioning, and seasonal and diurnal variability in the EC tower datasets. The results of the experiment also yielded insight on the practices followed by scientists to specify data properties, identified new data properties challenges, and proposed a method to capture data quality control confidence levels.

Keywords: Quality control processing, Sensor Data, Data analysis.

12.LSID versus HTTP URI: Two Approaches and E-Infrastructures for Managing Information about Taxon Names

Nina Laurenne[1], Jouni Tuominen[1], Arto Mertaniemi[2], Hannu Saarenmaa[3], Eero Hyvönen[1]
[1]Aalto University School of Science and University of Helsinki, [2]University of Helsinki, [3]University of Eastern Finland
nina.laurenne@helsinki.fi, jouni.tuominen@aalto.fi, arto.mertaniemi@helsinki.fi, hannu.saarenmaa@uef.fi, eero.hyvonen@tkk.fi

The amount of biological information has increased during the last decades. The information is hidden in museum specimens, data bases of observations, and in literature. Integrating data from scattered sources is hard because different vocabularies are used. The biggest barrier for data integration is the changing nature of scientific names, which hinders the interoperability of information systems. A solution is to use machine-processable identifiers for identifying biological names.

Two approaches are presented for managing taxon names. The first one is a taxonomic database of the Finnish Museum of Natural History based on Life Scinece Identifiers (LSIDs). The scientific names of six butterfly checklists are cross-mapped and linked taxon names form a concept to which an LSID is given. The concept covers a currently valid name, synonyms, and their lexical variants, and references to original publications and the year of publication if available. A tool for mapping taxon names between checklists is provided.

The second approach is based on HTTP URIs and the taxonomic metaontology TaxMeOn is presented for depicting the information of the butterfly species lists. The meta-ontology is based on RDF/OWL and the key classes are: a scientific name, a taxonomic concept, a name status, a taxonomic rank, a reference to a publication, an author and a common name. The same relations are applied for mapping taxon names as using LSIDs. TaxMeOn provides functionalities for humans and machines for accessing the ontologies that are published in the ONKI Ontology Service. ONKI supports content indexing, concept disambiguation, searching, and query expansion.

Cross-linking taxon names between species lists helps users to piece together the changes of scientific names and estimate the approximate amount of taxonomic treatments (none vs. many). Linking taxon names at the species level is straightforward, but at higher levels the problem is how to reconcile differing classifications of checklists.

The choice of an identifier used depends on needs, but despite the chosen identifier, the problem remains the same i.e. how to describe taxonomic information consistently without losing practicality. There is no significant difference whether LSIDs or HTTP URIs are used use to identify scientific names of checklists. However, the latter is more flexible as it allows interlinking the data with other Linked Data datasets increasing their interoperability.

Keywords: scientific name, identifier, data integration, species list, ontology

13.A Controlled Vocabulary for LTER Data Keywords

John Porter[1], Margaret O'Brien[2], Duane Costa[3], Donald Henshaw[4], Corinna Gries[5], Eda Melendez[6], Kristin Vanderbilt[7], Jason Downing[8], James Laundre[9]
[1]University of Virginia, [2]University of California, Santa Barbara, [3]LTER Network Office, Univ. of New Mexico, [4]USFS PNW Research Station, [5]University of Wisconsin, Madison , [6]University of Puerto Rico, [7]University of New Mexico, [8]University of Alaska, Fairbanks, [9]Marine Biological Laboratory
jporter@virginia.edu, mobrien@lternet.edu, dcosta@lternet.edu, dhenshaw@fs.fed.us, cgries@wisc.edu, emelendez@lternet.edu, kvanderbilt@lternet.edu, jdowning@lternet.edu, jlaundre@lternet.edu

The Controlled Vocabulary Working Group of the U.S. LTER Information Management Committee has completed work on a controlled vocabulary for science keywords and organized those keywords into a polytaxonomy for use in enhanced searching and browsing of data. The working group also developed a set of web-service-based tools for extracting terms, their synonyms, their narrower (child) terms, related terms and the narrower terms of those related terms. Additionally, to aid in the creation of metadata using the list, the working group developed auto-complete tools for web forms used to create metadata and tools that scan existing metadata and suggest suitable keywords. The controlled vocabulary incorporates over 600 keywords used at two or more LTER sites, or used by the National Biological Information Infrastructure Thesaurus, along with widely used synonyms. Structuring of the keywords into a polytaxonomy follows recommendations the NISO Z39.19 2010 standard, and the resulting structures are stored in a web-accessible TemaTres database (http://vocab.lternet.edu), which supports a variety of web services and is capable of exporting the structure in a variety of forms, including SKOS. The polytaxonomy has been incorporated in the the LTER Data Portal so that a search for “CO2” will automatically find datasets tagged with “carbon dioxide” as well as those simply tagged with “CO2” and a

search on "disturbance" will also return datasets tagged with common disturbances, such as floods or hurricanes. An advanced search capability also allows the user to select the level of enhancement applied to searches by controlling which types of related terms will be searched. Despite its simplicity, relative to more complex and semantically-rich structures (such as ontologies), the polytaxonomy has proven to be effective in increasing the reliability of searches for data on the LTER Data Portal. However, the working group has also begun the step of adding relationships to create a thesaurus that incorporates peer-to-peer links as well as parent-child relationships. This thesaurus can serve as a starting point for efforts aimed at developing additional semantic tools. The web services provided by TemaTres and developed by the working group make it relatively easy to enhance searching, browsing and keywording using a variety of interfaces.

Keywords: Keywords, Semantics, Polytaxonomy, Thesaurus

14. ESIP Federation: Using a Collaborative, Network Approach to improve Earth Science Interoperability

Erin Robinson, Carol Meyer
Foundation for Earth Science
erinrobinson@esipfed.org, carolbmeyer@esipfed.org

A variety of connections are needed across distributed communities in order to reach consensus on Earth science interoperability issues surrounding data discovery, access and quality. There is a paradigm shift currently happening away from silo-ed, monolithic systems, toward a loosely-coupled, networked, community-driven approach largely aimed at fostering Earth science interoperability between data, systems, people and organizations. The Federation of Earth Science Information Partners (ESIP Federation) fosters connections among a diverse community of practitioners across the Earth sciences and along the data value chain from data providers to application developers. Further, the ESIP Federation is fostering the development of a neutral research community that cuts across traditional discipline boundaries, enabling communities to share tools, data and technology. This synergy between cross-community collaboration, a commitment to openness, and broad practitioner expertise allows the ESIP Federation to play an important coordination role for the Earth science data and technology community. Ultimately, this coordination across sectors and communities will address problems central to the access and use of Earth science data and information, will allow Earth science research to be of higher quality and done more efficiently, and will leverage the work of the many communities contributing to Earth science knowledge. This poster will describe the strategic goals and activities the Federation is involved in, how we have used semantic web tools and methods to map our community and foster additional useful connections to further Earth science interoperability faster and will provide an invitation for others to leverage the ESIP collaboration space for their own Earth science interoperability needs.

Keywords: Community, Collaboration, Interoperabilit

15. The New Face of FLUXNET: Redesigning the Web Site and Data Organization to Enhance User Experience

Harold Shanafield[1], Ranjeet Devarakonda[1], Bob Cook[1], Stefanie Shamblin[1], Tammy Walker Beaty[1], Reid Boehm[2], Ben Mcmurry[1]
[1]Oak Ridge National Laboratory, [2]University of Tennessee
shanafieldha@ornl.gov, devarakondar@ornl.gov, cookrb@ornl.gov, hallsd@ornl.gov, beatytw@ornl.gov, mboehm@utk.edu, mcurrybf@ornl.gov

The FLUXNET global network of regional flux tower networks serves to coordinate the regional and global analysis of eddy covariance based CO2, water vapor and energy flux measurements taken at more than 500 sites in continuous long-term operation. The FLUXNET database presently contains information about the location, characteristics, and data availability of each of these sites. To facilitate the coordination and distribution of this information, we redesigned the underlying database and associated web site. We chose the PostgreSQL database as a platform based on its performance, stability and GIS extensions. PostreSQL allows us to enhance our search and presentation capabilities, which will in turn provide increased functionality for users seeking to understand the FLUXNETdata. The redesigned database will also significantly decrease the burden of managing such highly varied data. The website is being developed using the Drupal content management system, which provides many community-developed modules and a robust framework for

custom feature development. One of the features we are developing is a KML feed of tower sites. In parallel, we are working with the regional networks to ensure that the information in the FLUXNET database is identical to that in the regional networks. Going forward, we also plan to develop an automated way to synchronize information with the regional networks.

Keywords: FLUXNET, Drupal, Data archival, Sociology of collaboration and data sharing

16.GCE Data Toolbox: Metadata-driven Software for Data Acquisition, Quality Control and Synthesis

Wade Sheldon
GCE LTER
sheldon@uga.edu

The effort required to process, document, and quality control raw data from sensors is often a limiting step in bringing environmental data online. Similarly, the effort required to find, download and refactor data collected by others can prove limiting in large-scale synthesis efforts. However, the GCE Data Toolbox (MATLAB-based software developed at the Georgia Coastal Ecosystems LTER) has proven effective in overcoming both of these barriers. This software can automate processing of data collected by a wide variety of data logger systems, from initial acquisition through quality control and distribution of documented data sets and plots. It is equally adept at harvesting and integrating existing data from national monitoring programs and environmental databases (e.g. LTER ClimDB/HydroDB, USGS NWIS, NOAA NCDC, NOAA NERR). This poster provides a brief overview of this software, which is freely available under an open source license.

Keywords: software, quality, control, sensors, synthesis, MATLAB

17.Visualization Options for Environmental Scientists

Allison Smith[1], Susan Stafford[2] and Judith Bayard Cushing[1]
[1]The Evergreen State College, [2]University of Minnesota
smiall03@evergreen.du, stafford@umn.edu, judyc@evergreen.edu

Grand challenge environmental science data are complex, highly distributed, and heterogeneous, and span both time and space. Cross-scale analytical methods are not well understood, so visualizing natural phenomena could be used in preliminary studies to help scientists hone intuition and sharpen testable hypotheses, additionally facilitating communication to broad audiences. Recently funded by the National Science Foundation (BIO/ABI/DBI:1062572), the VISualization of Terrestrial-Aquatic Systems (VISTAS) project, a collaboration among The Evergreen State College, Oregon State University, and the HJ Andrews Long Term Ecological Research (LTER) site, aims to facilitate understanding and communication of ecological processes through visualizations of complex, heterogeneous data sets. Computer scientists and ecologists are collaborating to develop software to assist in visual analytics at scales. In addition to software development, the project will work with social scientists to study VISTAS' software and visualization co-development and assess the usability of VISTAS and its visual analytics, and to answer the critical question: Which visualizations work, for what purposes, with which audiences?

The initial phase of our research has involved a survey of LTER Information Managers to identify and critique visualization tools used by LTER scientists, and review of non-commercial visualization tools that integrate data sets for synthesis science. Information Managers were asked: "What tools are you and the scientists at your site using for visualization, map making, chart development, and other visual outputs?" They were then asked to assess the effectiveness of those tools. If the tools were not effective, they were asked: "What tools and capabilities are missing?" Our preliminary survey of open-source and "free" software and U.S. visualization centers focused on efforts to provide 2- and 3-D visualization.

This poster profiles currently available visualization software (noting specific capabilities) and articulate some areas of inquiry not currently addressed.

For further information, see http://blogs.evergreen.edu /vistas or contact Judy Cushing, judyc@evergreen.edu. We gratefully acknowledge 1) the work of VISTAS researchers in contributing insights into visualizing environmental data: Barbara Bond (HJAndrews LTER); Denise Lach, John Bolte (Oregon State University); Nik Stevenson-Molnar (Conservation Biology Institute), et al, and 2) the LTER Information Managers who responded to our request for information: John Chamblee, Inigo San Gil, Don Henshaw, Eda Melendez, Margaret O'Brien, John Porter, Linda Powell, Mark Servilla, Wade Sheldon, Theresa Valentine, and Kristin Vanderbilt.

Keywords: Data Visualization, Visual Analytics, Usability assessment

18. Web-based Visualization Tools for Remote Sensing Data

Suresh K Santhana Vannan, Robert B. Cook, Yaxing Wei, Chris W. Lenhardt
Oak Ridge National Laboratory
santhanavans@ornl.gov, cookrb@ornl, gov weiy@ornl.gov, lenhardtc@ornl.gov

Remote sensing data are highly useful for environmental and terrestrial ecology research. The diversity and availability of remote sensing data products have made them an important data source for analyzing key science questions relating to Earth System processes at regional, continental, and global scales. Remote sensing data are also useful to understand environmental characteristics (land cover, soil, vegetation) and dynamics (vegetation phenology and productivity). Remote sensing data products can be created, distributed, and used in diverse projections and formats as well. However, this diversity can hinder the usability of the data, and limit data users' abilities to visualize, interpret and use these data sets for science and application purposes.

To enhance the usability of remote sensing data products, the Oak Ridge National Laboratory (ORNL) Distributed Active Archive Center (DAAC), a NASA-sponsored terrestrial ecology data center, has used geospatial Web service standards and other Web-based visualization technology to increase the understanding, usability and availability of remote sensing data products. Through the ORNL DAAC visualization tools remote sensing data sets are standardized into non-proprietary file formats and distributed through custom Web tools supporting Open Geospatial Consortium (OGC) Web Service standards. In particular, ORNL DAAC has developed time series and grid visualization tools for Moderate Resolution Imaging Spectroradiometer (MODIS) sensor data products. Web tools and Web services provide MODIS subsets in comma-delimited text format and in GIS compatible GeoTIFF format. Users can download and visualize MODIS subsets for a set of pre-defined locations or order/visualize MODIS subsets for any land location. Web based GIS tools are also available to visualize the MODIS subsets. The ORNL DAAC has also created a Web-based application called Spatial Data Access Tool (SDAT) that is based on OGC Web services standards and allows data distribution and consumption for users not familiar with OGC standards. SDAT also allows for users to visualize the data set prior to download. Google Earth visualizations of the data set are also provided through SDAT. Remote sensing data products such as Advanced Land Observing Satellite (ALOS) - PALSAR (Phased Array type L-band Synthetic Aperture Radar) Synthetic Aperture Radar subsets, Landsat, Shuttle Radar Topography Mission (SRTM) and MODIS land cover data are also available through the SDAT visualization tool. This poster provides description of the ORNL DAAC visualization tools with technical details, access information, and use case examples.

Keywords: Remote Sensing, Visualization, Web Standards, Interoperability

19. The CENS Data Registry: An Evolving Approach to Data Repositories for Scientific Data Discovery and Reuse

Jillian Wallis, Christine Borgman, Matthew Mayernik and Alberto Pepe
University of California, Los Angeles
jwallisi@ucla.edu, borgman@gseis.ucla.edu, mattmayernik@ucla.edu, apepe@ucla.edu

Science and technology research is becoming not only more distributed and collaborative, but more highly instrumented. Data repositories provide a means to capture, manage, and access the data deluge that results from these research enterprises. We have conducted research on data practices and participated in developing data management services for the Center for Embedded Networked Sensing (CENS) since its founding in 2002 as a multi-institution, multi-domain National Science Foundation Science and Technology Center. Over the course of eight years, our data repository strategy has shifted dramatically in response to changing technologies, practices, and policies. As CENS has evolved, so has the larger social and political framework for data sharing. Now that we have external pressure to share our data, we have more mechanisms to establish policies and systems. Our approach to sharing data has come full circle from building a data repository to building a metadata repository that will enable CENS data to be discovered. Once discovered, prospective users can obtain data from wherever they may be located, and from whoever has the rights and ability to release them. CENS scientific and technological activities have evolved at Internet speed over the course of eight years. The focus shifted from long-term, static deployments to short-term, dynamic campaigns, and from single-purpose sensing technologies to cell phones as mobile sensing devices. Concurrently, the Internet has moved from basic web services to "web 2.0" and cloud computing. Data repository technologies and services have not evolved as rapidly as the applications they serve. We've worked

intensively to keep pace with these moving targets over the last eight years, and the pace shows no signs of slowing. In this poster we report on the development of several data repository systems and on the lessons learned, which include the difficulty of anticipating data requirements from nascent technologies, building systems for highly diverse work practices and data types, the need to bind together multiple single-purpose systems, the lack of incentives to manage and share data, the complementary nature of research and development in understanding practices, and sustainability.

Keywords: distributed research, collaborative research, data practices, data repositories

20. Interoperable Geospatial Data for Carbon Cycle Research

Yaxing Wei, Robert Cook, Wilfred Post, Jerry Pan, Chris Lenhardt
Oak Ridge National Laboratory
weiy@ornl.gov, cookrb@ornl.gov, wmp@ornl.gov, pany@ornl.gov, lenhardtc@ornl.gov

Carbon cycle research is data intensive. One major barrier is that data, especially geospatial data, involved in carbon cycle research is usually heterogeneous, distributed across multiple organizations, and lacking the mechanisms to be easily shared by a broad user community.

In Oak Ridge National Laboratory (ORNL), the NASA-sponsored Modeling and Synthesis Thematic Data Center (MAST-DC) promotes carbon cycle research by leveraging emerging standards to prepare standardized geospatial data that can be discovered, accessed, understood, and used by carbon cycle researchers and tools in an easy and interoperable way.

The MAST-DC supports North American Carbon Program (NACP) multi-scale synthesis and terrestrial model inter-comparison and evaluation activities by processing, managing, and distributing carbon cycle model input and output data, observational, and inventory data products. The MAST-DC compiles a variety of environmental driver data sets, including climate, soil, vegetation, biome classification, land use change, and nitrogen deposition, into Climate & Forecast (CF) convention compatible netCDF format. It also standardizes output data from more than 20 terrestrial biospheric models into CF-compatible netCDF format with common attributes, including spatial/temporal resolution and extent, units, and coordinate reference system. Related observational and inventory data sets, including MODIS GPP/NPP and forest inventory estimates, are also prepared in CF-compatible netCDF format. The standardization process involves two aspects: the standardization of data content and standardization of metadata. Interpretation metadata is prepared by following CF-1 convention and FGDC standard is used for the representation of discovery metadata. These standardized data are then fed into a well-designed Spatial Data Infrastructure (SDI). Open Geospatial Consortium (OGC) standards-based Web Coverage Service (WCS) and Web Map Service (WMS) and OPeNDAP service are utilized to provide on-demand visualization and access to geospatial data for carbon cycle researchers. These data sets are also made discoverable in an interoperable way by feeding FGDC-compatible metadata into an OGC Catalog Service for Web (CSW) service.

It has proved that this standards-based approach increases the interoperability of geospatial data and benefits both the modeling teams and the researchers performing synthesis and model evaluation activities. Additional experiments will be carried to investigate how the data products in the MAST-DC can interoperate with existing cyber-infrastructures involved with carbon cycle research, including Earth System Grid (ESG) and Data Observation Network for Earth (DataONE).

Keywords: data, synthesis, interoperable, geospatial, carbon cycle, ogc, netcdf