



## **FINAL REPORT**

### **2011 Workshop on Aquatic Ecosystem Sustainability (WAES)**

**June 13-14, 2011**

**Marina del Rey, CA**

<http://water.isi.edu/waes11>

**Yolanda Gil and Thomas Harmon (Eds)**

Technical Report No. ISI-TR-674  
Information Sciences Institute  
University of Southern California

*This report was written by Matt Becker, Dan Crichton, Ewa Deelman, Yolanda Gil, Tom Harmon, Stephanie Granger, Qinghua Guo, Paul Hanson, Andreas Hofmann, Burt Jones, Craig Knoblock, Mike McCann, Timothy Stough, Pedro Szekely, Ryan Utz, and Sandra Villamizar.*

*Photo credits: Jason Riesa.*

**ISI Technical Report Number ISI-TR-674**

**October, 2011**

Information Sciences Institute  
University of Southern California

4676 Admiralty Way  
Marina del Rey, CA 90292

<http://www.isi.edu>

## Table of Contents

<b>TABLE OF CONTENTS</b> .....	<b>3</b>
<b>WORKSHOP PARTICIPANTS</b> .....	<b>4</b>
<b>SPONSORS</b> .....	<b>4</b>
<b>EXECUTIVE SUMMARY</b> .....	<b>5</b>
<b>CHALLENGES</b> .....	<b>5</b>
<b>RECOMMENDATIONS</b> .....	<b>6</b>
<b>1 INTRODUCTION</b> .....	<b>7</b>
<b>2 MOTIVATING SCENARIOS</b> .....	<b>9</b>
2.1 RIVERS.....	9
2.2 LAKES.....	10
2.3 OCEANS.....	11
<b>3 CHALLENGES IN CURRENT PROCESSES</b> .....	<b>11</b>
3.1 USING DATA FROM COMMUNITY REPOSITORIES.....	11
3.2 PROCESSING DATA ACROSS TOOLS.....	13
3.3 PUBLISHING DATA AND PUBLISHING MODELS .....	14
<b>4 COMMON THEMES AND RELEVANT TECHNOLOGIES: NEW OPPORTUNITIES</b> .....	<b>15</b>
4.1 DATA CLEANING AND NORMALIZATION .....	15
4.2 WORKFLOWS.....	16
4.3 METADATA AND OPEN DATA PUBLICATION ON THE WEB .....	17
4.4 PROVENANCE-AWARE TOOLS.....	19
<b>5 OBSERVATIONS AND INSIGHTS</b> .....	<b>20</b>
5.1 COMMUNITY NEEDS.....	20
5.2 PERCEIVED TECHNOLOGY NEEDS.....	21
5.3 COLLABORATION CHALLENGES.....	22
<b>6 VISION: ENABLING NEW AQUATIC ECOSYSTEM SCIENCE</b> .....	<b>23</b>
6.1 ENABLING NEW AQUATIC ECOSYSTEM SCIENCE: NITROGEN AND CARBON DYNAMICS.....	23
6.2 ENABLING TECHNOLOGIES: ENVISIONING A “DATA LIBRARIAN” .....	26
<b>7 CONCLUSIONS</b> .....	<b>28</b>
<b>8 REFERENCES</b> .....	<b>29</b>

## Workshop Participants

[Matt Becker](#), Department of Geological Sciences, California State University Long Beach

[Terry Benzel](#), Information Sciences Institute, University of Southern California

[Amy Braverman](#), Jet Propulsion Laboratory (JPL)

[Dan Crichton](#), Earth Science and Technology, Jet Propulsion Laboratory (JPL)

[Todd Crowl](#), Department of Watershed Sciences, Utah State University

[Ewa Deelman](#), Information Sciences Institute, University of Southern California

[Yolanda Gil](#), Information Sciences Institute, University of Southern California

[Tom Harmon](#), School of Engineering, University of California Merced

[Stephanie Granger](#), Jet Propulsion Laboratory (JPL)

[Qinghua Guo](#), School of Engineering, University of California Merced

[Paul Hanson](#), Center for Limnology, University of Wisconsin at Madison

[Andreas Hofmann](#), Monterey Bay Aquarium Research Institute (MBARI)

[Burt Jones](#), Biology Department, University of Southern California

[Craig Knoblock](#), Information Sciences Institute, University of Southern California

[Mike McCann](#), Monterey Bay Aquarium Research Institute (MBARI)

[Timothy Stough](#), Jet Propulsion Laboratory (JPL)

[Pedro Szekely](#), Information Sciences Institute, University of Southern California

[Ryan Utz](#), National Ecological Observatory Network (NEON), Aquatic/STREON

[Sandra Villamizar](#), School of Engineering, University of California Merced

## Sponsors

The workshop was sponsored by the [Information Sciences Institute](#) of the University of Southern California and by the [School of Engineering](#) of the University of California Merced.

The University of Southern California's Information Sciences Institute (ISI) is a world leader in many areas of computer science such as computer networks, distributed systems, and artificial intelligence. Part of USC's Viterbi School of Engineering, ISI is known for excellence in basic and applied research.

UC Merced is the first new American research university in the 21st century, with a mission of research, teaching and service.

## Executive Summary

Environmental cyberobservatory (ECO) planning and implementation has been ongoing for more than a decade now, and several major efforts have recently come online or will soon. Some investigators in the relevant research communities will use ECO data, traditionally by developing their own client-side services to acquire data and then manually create custom tools to integrate and analyze it. However, a significant portion of the aquatic ecosystem science community will need more custom services to manage locally collected data. The latter group represents enormous intellectual capacity when one envisions thousands of ecosystems scientists supplementing ECO baseline data by sharing their own locally intensive observational efforts.

The Workshop for Aquatic Ecosystem Sustainability (WAES) convened in June 2011 and focused on the needs of aquatic ecosystem research on inland waters and oceans. A key observation is that integrative and holistic research results are needed to address the sustainability of these interconnected life-supporting ecosystems. Only in this manner can we gain an adequate understanding of the impact of changing environmental conditions due to climatic variation and ongoing or future human activities in adjacent watershed on the biodiversity and ecosystem functions offered by aquatic ecosystems.

## Challenges

Workshop participants agreed that the sustainability of these ecosystems depends on the ability of the research community to: 1) readily access data with local records that is collected and owned by individual scientists and that have bearing in global phenomena, 2) easily integrate local scientist data with existing data catalogs made available by cyberobservatories, government institutions, and non-governmental organizations, and 3) efficiently process data and model phenomena in a repeatable and understandable manner.

Workshop participants agreed to several observations about the challenges faced by the community:

- Challenges and technology requirements are common across many research areas in environmental sciences.
- There are many datasets collected and owned by individual scientists in a variety of areas in environmental and ecosystems sciences that are extremely valuable and yet are not shared.
- Scientists collect datasets separately and maintain them locally, investing enormous amounts of time organizing and preparing data that, while slightly altered in format, are similar in nature.
- The difficulty of finding scientists' local datasets is so great that many opportunities for regional and global synthesis are in danger of being lost.
- Many environmental scientists are unaware of relevant advances in computer science, particularly in rapidly changing areas that are not traditionally connected with environmental sciences.
- Workflows have been defined in several communities and are used for explicit process sharing with a number of benefits.
- Environmental science would benefit for many reasons from faster turnaround from sensor to analysis.
- The continuity of provenance and other metadata improve the usefulness of the data to individual scientists and enable the reuse of the transformed data by other scientists.
- Remote sensing is positioned to have an immediate impact in environmental sciences if its use were better supported by infrastructure.
- Facilitating the creation of shared data formats and metadata properties would be very beneficial.

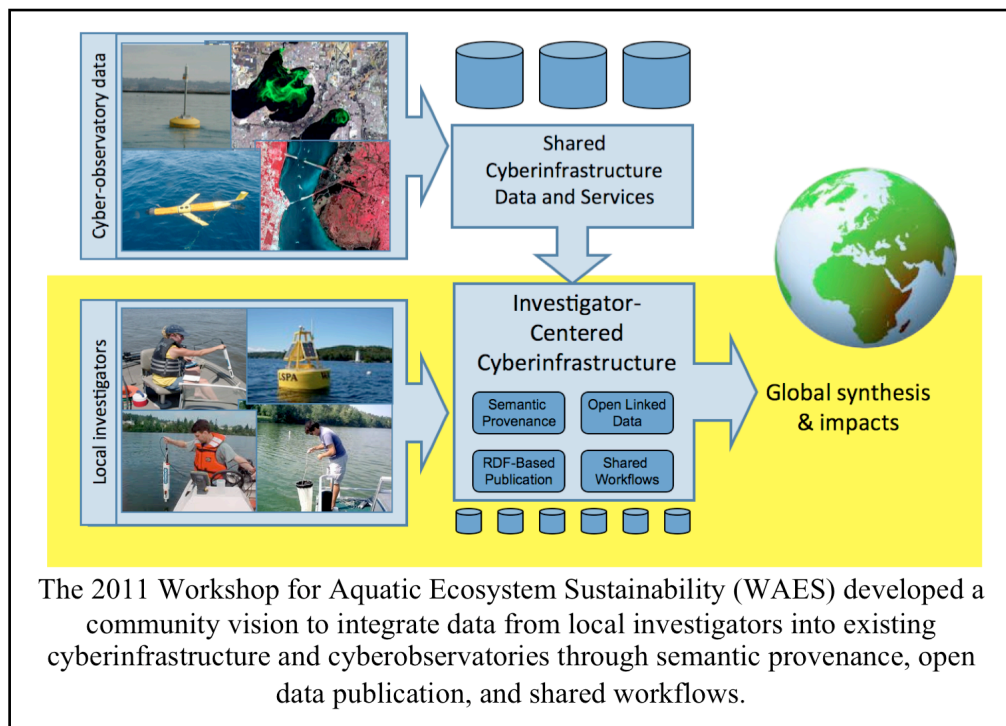
## Recommendations

Workshop participants advocated new approaches to support scientists in the analysis, integration, and modeling of data based on:

- A new breed of software tools in which semantic provenance is automatically created and used by the system,
- The use of open standards based on RDF and Linked Data Principles to facilitate sharing of data and provenance annotations,
- The use of workflows to represent explicitly all data preparation, integration, and processing steps in a way that is automatically repeatable.

The key recommendations from the workshop participants are:

- Data sharing approaches that reduce publication cost and provide immediate benefits to the scientist are important in order to rescue abundant and likely insightful data in environmental and ecosystems science that will otherwise be lost.
- Workflow systems that manage and automate data processing steps are crucial to enable efficient processing of the volumes of data required to address regional, continental, and global-scale environmental and ecosystems science questions.
- More efficient processing of environmental and ecosystems science data would improve data collection by enabling rapid adjustment of sampling and sensor configurations.
- Pervasive provenance recording would improve reuse and productivity by facilitating the flow of data and processes across research groups and disciplines.
- Workflow sharing can drive and facilitate collaborative science projects that represent higher-impact science efforts.
- Data and software repositories should exhibit an economy of scale where individual efforts save effort to others.



# 1 Introduction

The transformation within the ecological community from broadly distributed independent investigators to a more team approach to science [Olson et al 2008], coupled with the recent and rapid emergence of ecological sensor networks [Porter et al. 2009], demands more standardized and organized ways of coupling information management with the research process. As a field science, ecology traditionally associates a data set with a PIs laboratory or even with the individual scientist in many cases. This is in contrast with physics, astronomy, and climate modeling, for example, where there are major group projects resulting in large shared data sets collected in concert [Graham et al 2008].

There are hundreds of field stations within the US alone (Organization of Biological Field Stations) and thousands of field scientists (Ecological Society of America). Yet, field programs in ecology typically do not have the resources or technology to develop their own information management systems and make them compatible with major national initiatives. Often data are sequestered or even lost after a focused research project is completed, wasting valuable information regarding long term ecologic or hydrologic trends. There is a need to enable a new kind of participatory science that leverages efforts from the individual laboratory toward achieving a common long-term goal at the national or even global scale. However, enabling

The benefits of cyberobservatories to the scientific community follows what is known as a long tail distribution, i.e., relatively few benefit tremendously, but a very large community of scientists benefits very little if at all. This long tail of scientists are in possession of troves of precious environmental data that they collect locally and that are very often not shared. For them, the cost of data sharing is prohibitive, while the benefits are unclear. There is a need to enable a new kind of participatory science that leverages efforts from the individual laboratory toward achieving a common long-term goal at the national or even global scale.

that participation is a major challenge, due in part to the many and varied types of data that are collected, how they are annotated (metadata), organized (ontology), and stored, and how disparate data sets are brought together in a common analytical framework. Overcoming these challenges are becoming increasingly difficult every year, as the advancement in sensor and network technology outpaces the information management systems and the scientists capable of handling the data streams.

Environmental cyberinfrastructure and cyberobservatory (ECO) efforts have been ongoing for more than a decade [OOI 2005; Keller 2008]. ECO developments by the Ocean Observatory Initiative [OOI 2010], and by the ecological [Keller 2008; Hanson 2008] and hydrologic science communities [WATERS 2009; CZO 2010] are evidence that the current suite of sensing technologies is adequate to launch these observatories. Realizing the full potential of recently established observational networks created to allow large spatiotemporal-scale ecological inquiry will require highly effective transmission of complex data to scientific communities. Fortunately, many such institutions have recognized the value of partnering with computer scientists to aid in facilitating data access, coordination, and quality. For example, the National Ecological Observatory Network (NEON) will collect, store and serve 30 years of open-access biological, chemical, geophysical and atmospheric data using standardized methods throughout North America to allow ecological forecasting at the continental scale [Keller et al. 2009, NEON 2009]. Data products from NEON aquatic ecosystems will include >200 base-level (i.e. NO<sub>3</sub> loads, discharge and algal chlorophyll concentrations) and approximately 15 high-level (such as benthic macroinvertebrate diversity and nitrogen flux) parameters collected in 36 streams, lakes and rivers [Keller 2010, Keller et al. 2010]. The aquatic site network is designed to allow continental-scale ecological forecasting and provide comprehensive site-specific data that may complement and enhance local agency- or principle investigator-led research. All NEON data will be made publicly available through an open-access web interface [Beasley 2010]

to be developed by an in-house cyber infrastructure team. The data discovery and acquisition system design will attempt to avoid the problems of past initiatives: data-seekers will be able to use search or filter functions to obtain data subsets, choose among several data output formats, and metadata standards (including information on provenance) will meet or exceed those employed by federal agencies. The cyber infrastructure system will use dataset provisioning standards, such as EML [EML 2011], NetCDF [NetCDF 2011], and output formats in attempt to maximize data compatibility with research efforts independent of NEON.

That massive new data streams are coming online is not in question. Environmental COs coming online are multi-scale sensor systems, including (1) stationary in situ sensor networks or webs monitoring continuously in time and deployed at various spatial scales, (2) mobile sensors (e.g., installed on autonomous underwater vehicles, AUVs) periodically mapping spatially distributed properties between stations, (3) aircraft- or satellite-based remote sensing products, and, more recently, (4) participatory (human-assisted) data acquired using microcomputers such as smart phones or PDAs. And even outside the auspices of these major observatory efforts, many individual investigators are actively employing one or more of these technologies to support their personal research portfolio. In both cases, increasingly large data volumes are accumulating, enabling investigators to focus more on science questions and less on separating key observations from variable environmental conditions.

However, as CO data continues to come online, an overarching question in many minds concerns whether or not the relevant scientific communities are in a position to rapidly deliver the intended scientific return of these large-scale investments.

The workshop on Workflows for Aquatic Ecosystems Sustainability (WAES 2011) convened an interdisciplinary group to examine the following question:

*How do the relevant research communities guarantee the meaningful data input, mining, analysis and synthesis by individual scientists (i.e., members of the “long tail”) needed to ensure that rapid scientific and policy-related impacts stem from environmental observatories?*

The premise for this question is that while concentrated community intellectual resources are being expended on ECO research and development, there remains a major resource of relatively untapped scientific bandwidth from researchers operating on individual or small group bases. That is, the benefits of ECO to the scientific community follows what is known as a *long tail distribution* [Anderson 2004]: relatively few benefit tremendously, but a very large community of scientists benefits very little if at all. Conversely, this long tail of scientists are in possession of troves of precious environmental data that they collect locally and that are not integrated into ECO infrastructure or very often not shared with the scientific community. For them, the cost of participating in ECO efforts is prohibitive, while the benefits are unclear and definitely not immediate. WAES workshop participants discussed the need to tap this resource, both in terms of integrating long tail data with the ECO data and enticing long tail scientists to incorporate ECO data into their investigation.

The WAES workshop participants examined the key bottlenecks and challenges to engage a greater portion of the scientific communities in the use of ECO, ancillary, and legacy data streams to study aquatic ecosystems. Specific outcomes from this WAES workshop stemmed from the consensus that a large portion of individual scientists are in dire need of help in terms of adding and leveraging off of ECO capabilities. Even as scientists grow increasingly comfortable with today’s technology, they typically carry out many manual steps of the process of configuring instruments to collect data, cleaning and archiving the data, and configuring, and calibrating and executing the models needed to synthesize the complex information, enabling them to test their hypotheses. These individual scientists also need help finding additional existing data (including legacy data and newer data



streams from the COs) that could support their research, implying a great need for sound and accessible metadata and provenance practices from the perspective of the CO data and the ancillary data flowing from the long tail scientists.

To address these challenges, new approaches are needed that complement existing ECO investments and extend them with infrastructure to support scientist-centered processes that will help to accelerate progress and enable new discoveries in environmental sciences.

This document begins proposing science questions in aquatic ecosystem research that require new infrastructure for scientist-centered processes, showing scenarios for rivers, lakes, and ocean systems. It then outlines the challenges involved in improving current practices for data sharing, data reuse, and data processing. Recent technologies directly relevant to those challenges are discussed. A synthesis of science needs and their correspondence to technology requirements is laid out as a research agenda for a vision of what new science would be enabled by broadening participation of the long tail of scientists.

## 2 Motivating Scenarios

Aquatic ecosystem research spans throughout communities that focus on rivers, lakes, coastal areas, or oceans. The scientific challenges faced expose significant common underlying threads.

### 2.1 Rivers

The synergistic growth of science and technology has made it possible for researchers to think about large-scale spatio-temporal processes which require a strategy that combines holistic and reductionist approaches for a positive advancement [Newman *et al.* 2006]. Among the major challenges within river systems are: (1) Integrating processes among scales such as climate change and flow regime within a watershed; (2) Understanding and predicting the response of species to the changing environmental conditions and the determination of thresholds that could trigger physiological or behavioral changes [Woodward *et al.* 2010]; and (3) Monitoring, modeling and predicting the transport and fate of contaminants across the different pathways within the watershed [Capel *et al.* 2008]. This comprehensive approach is critical not only for the advancement of science, but for the implementation of policies and regulations that respond to the particular necessities of the system.

The common aspect of all these scientific questions is that their assessment requires the use and manipulation of a widespread variety of data sources. Academic or governmental and non-governmental institutional data obtained either through web services, other digital formats, or paper records including: field observations (e.g., river stage, fish surveys), derived parameters (e.g., rating curves), laboratory results (e.g., specific nutrient concentrations), river model inputs and outputs (e.g, discharge and stage forecasts). The heterogeneous nature of data sources hinders the ability of the scientists to focus on answering their particular research questions. Consider, for example, river discharge, a primary hydrologic variable and the dominant control variable in lotic ecosystems [Poff *et al.*, 1997]. Discharge records abound in open-access data storehouses and

Participatory science challenges include: 1) Integrating processes among scales such as climate change and flow regime within a watershed; 2) Understanding and predicting the response of species to the changing environmental conditions and the determination of thresholds that could trigger physiological or behavioral changes; and 3) Monitoring, modeling and predicting the transport and fate of contaminants across the different pathways within the watershed. A comprehensive approach is critical not only for the advancement of science, but for the implementation of policies and regulations.

represent a wealth of research opportunities that could be greatly facilitated with improved data access systems. Fortunately, many governmental agencies maintain long-standing programs that collect high-resolution temporal discharge data, resulting in records that in some cases date back to the nineteenth century. Hydrologists have long assessed these data to detect important spatiotemporal patterns in snowmelt timing [Stewart *et al.* 2005] related to climate change, heterogeneity in flow regime structure among geoclimatic regions (e.g., [Poff *et al.* 2006]), and variability in how land use change impacts ecosystems [Utz *et al.* 2011]. In spite of the fact that these data are typically open-access, and tools such as Web services are becoming available for accessing many relatively current data streams [Maidment 2008; USGS 2011], amalgamating records from multiple sites or at fine resolutions often proves challenging and prohibitively time-consuming.

As the preceding example suggests, systems that streamline multisite data recovery and/or assimilate watershed data and metadata within and among agencies would greatly enhance efforts to assess patterns in flow regimes (and therefore dependent ecosystem processes and functions) at coarse spatial and temporal scales. In such cases, where the measured and derived parameters have already been standardized, we see a need for greater assistance to the scientist by automating common sets of procedures for data acquisition, data analysis, and ultimately synthesis of the resulting homogenized data streams.

## 2.2 Lakes

The study of lake ecosystems tends to be geographically limited, leading to a somewhat balkanized community of research. GLEON is an international grassroots network of ecologists, physical limnologists, and information technology experts who have a common goal of building a scalable, persistent network of lake ecology observatories in order to improve understanding and management

In scientific collaborations there remains a fundamental bottleneck in transforming tractable ideas into scientific products. That bottleneck is accepting data contributions from disparate data sources and preparing those data for use in a common analytical framework.

of lake ecosystems. The Global Lake Ecological Observatory Network (GLEON) provides several examples of how network science demands a highly organized approach to assembling data sets from around the world [Hanson 2007]. Working groups within GLEON form around scientific themes, such as microbial ecology, lake physics and climate change, metabolism, and phytoplankton ecology. The kinds of

scientific questions that emerge from working groups inevitably call for contributions of data sets from lake observatories (or individuals). For example, the question, “What controls ecosystem respiration in lakes?” spawned an effort to collect high-frequency sensor data from nearly 30 lake observatories. The question, “What are the controls over phytoplankton assembly?” has led to the contribution of long-term phytoplankton data sets to the working group. Both of these efforts have resulted in extremely valuable data sets that have spawned additional research questions and projects.

However, the lag between formulating an important and answerable science question and realizing results imposes many logistical challenges, especially for organizations that operate virtually (i.e., the people and resources are broadly distributed), as GLEON often does. In scientific collaborations there remains a fundamental bottleneck in transforming tractable ideas into scientific products. That bottleneck is accepting data contributions from disparate data sources and preparing those data for use in a common analytical framework. There are many steps involved, including: (a) agreeing on a controlled vocabulary and minimal metadata so that all can understand the contents of contributed data sets; (b) defining a set of acceptable structures for data sets so that software tools can load and parse data; (c) performing basic QA/QC on data to eliminate outliers and identify suspect values; (d)

fill gaps in data, depending on whether the size of the gap is incidental or of potential consequence for the analysis; (e) synchronize variable vectors in space or time so analysis software can consume the data; (f) visualize data. These steps make up a common process for data analysis and can be reasonably well defined in a general sense for lake aquatic ecosystems. In spite of the commonality of this process, each research project tends to re-invent the steps in an ad hoc basis.

## 2.3 Oceans

Ocean ecosystem research involves gaining a better understanding of the transfer of matter and life in all areas of the ocean. Improved understanding will lead to better predictive capabilities for global-scale processes, including climate change, and for local-scale events, such as toxic algae blooms and low-oxygen dead zones. Advances in robotics, sensor technology, and global satellite data communications over the last few decades permit orders of magnitude more *in situ* observations than were previously possible with traditional ship-going techniques. Achieving maximal use of these new observations presents a challenge for the collection of data systems (and people) that are tasked with managing these important archives. Though the oceanographic community has recognized this issue [OceanObs 2009] and has a blueprint for employing data standards (e.g., NOAA IOOS [IOOS-DMAC 2010; de La Beaujardiere 2008]), there remain significant challenges.

The ocean science community continues to operate with technical personnel that can be euphemistically described as “pretty good at Matlab™”. For the typical researcher this provides the shortest path from collecting the observations, processing and archiving the data, and analyzing the data for publication. Unfortunately, this operational paradigm results in scientific output that is unrepeatable and obtained from original data that is unavailable to the larger scientific community. By streamlining data analysis processes within the community we can promote a new paradigm where data are routinely connected

The current operational paradigm results in scientific output that is unrepeatable and obtained from original data that is unavailable to the larger scientific community. By streamlining data analysis processes within the community we can promote a new paradigm where data are routinely connected from the particulars of the instrument deployment to the resultant scientific product.

from the particulars of the instrument deployment to the resultant scientific product. This includes: 1) Tooling for provenance capture within applications that scientists use; 2) Coupling of the metadata to the data stream from the collection phase through finished products, analogous to photo image Exif metadata; 3) Software and data engineering training for those who process data; 4) Publication policies that encourage publication of analysis processes and software, not just data.

## 3 Challenges in Current Processes

Participants discussed the challenges in processing environmental data through a variety of tools and modes, integrating individually collected data with data shared repositories, and publishing data to enable reuse by others.

### 3.1 Using Data from Community Repositories

Many community repositories are available to environmental scientists. We mentioned NEON and OOI in Section 1. There are also government run data sources, some at the local and regional levels. Non-governmental organizations also collect environmental data that is made publicly available. In

addition, agencies such as NOAA and NASA play a major role in providing infrastructure for collecting and managing remote sensing data.

Significant work is still required to deal with the heterogeneity and to develop the infrastructure that allows for data across shared repositories to be brought together and disseminated in more dynamic ways.

Shared scientific data repositories have long focused on the critical needs to capture data from upstream data producers (instruments, sensors and other scientific activities). However, what is becoming important in addressing the needs of the scientific community is turning these repositories into useful knowledge bases. Many disciplines are constructing “pipelines” that provide

processing and generation of scientific data sets that are ultimately housed in these data repositories. Within Earth Science remote sensing, for example, pipelines that produce data higher order data products (calibrated, gridded, etc) are generally included as part of the ground system in when new observing satellites are launched. Downstream, however, many scientists are pulling data from these repositories, reprocessing and reformatting data, and then including them in their experiments. This may include fusing data from multiple sources, observations and platforms. This begs the question of whether improvements in this process can be achieved to drive more dynamic scientific discovery capabilities from online data repositories.

While many of these pipelines are now traditionally constructed for each instrument or mission, there is an increasing interest in performing analysis across data sets that may span different instruments or missions, even disciplines. This type of data integration and intercomparison is not bound to just observational data sets. Within the climate research community, effort is underway to prepare observational data so that it can be compared against climate models. However, within climate as with other disciplines, data is captured within institutions, systems and structures using different standards and measurements that make such analysis difficult due to the heterogeneity. The overarching need is to ensure these repositories have services that provide scientists the ability to directly integrate them to enable more dynamic analysis.

The movement towards building services around data repositories is critical to enabling more interdisciplinary science and improving the dynamic discovery capabilities. Traditionally, scientists have developed their own client-side analysis environments, pulling down data from each repository. As part of the process, significant work is performed to reformat and transform the data to enable it to be combined and/or compared against data from other repositories. As a result, these processes are tightly coupled on the client-side and do little to dynamically integrate remote repositories. Rather than following the path of integrating scientific workflows into client-side scientific applications developed by each investigators, the emphasis should be on developing services that support query, discovery, subsetting, reformatting, regridding and other basic analytics that allows for generating workflows that combine and process data from distributed sources. This is a shift in the paradigm which puts more emphasis on getting the data into repositories to begin the data processing functions as part of the science investigation vs. setting up rigid pipelines. This shift also allows for construction of new computational services that can be provided and allow for growth in capabilities over time.

In addition to having appropriate services, another key need in generating scientific data for these repositories, is capturing both the raw and processed data along with high quality metadata to annotate it. A particularly critical piece of that metadata is the provenance information. This includes capturing the detailed steps and being able to repeat these steps are critical to ensuring that scientific results can be validated and it is essential that any processed data capture the provenance information as part of its core metadata. This includes information about the origin, versioning, processing history, decisions made that impacted generation of the data, etc. In addition, having

consistent, reliable metadata is important to support effective search, retrieval and fusion of these data. All too often, data repositories used their own internal discipline information models that annotate the data, but do little to support federation of data across multiple repositories since their underlying data models differ.

In summary, many scientific communities are now recognizing the importance of developing both the scientific software services and the necessary information models that are needed to build higher quality repositories and to better annotate the data sets that are captured. However, significant work is still required to deal with the heterogeneity and to develop the infrastructure that allows for data across these repositories to be brought together and disseminated in more dynamic ways.

### 3.2 Processing Data Across Tools

With the collaboration across disciplines within environmental sciences, the need for standardizing data has become apparent in order to facilitate data sharing among different tools. Data includes temporal and spatial data as well as their associated meta-data and database structures used for data storage and retrieval.

Today, the raw data collected from sensors is processed through several tools before it becomes usable. A first tool, the sensor, is the raw data collection layer (Some sensors require more processing, but if the data can be described, it can be streamed. Otherwise, it needs to be described before streaming). That data is typically processed by another tool to describe and annotate it, as the translation layer in which the raw data is interpreted into something understandable (usable). Next, the data may be processed by a formatting tool, where described data is stored in a structured way to be saved and shared. Then, another tool for quality assurance/quality control (QA/QC) checks the data for completeness. At this point the data is considered ready for use. Throughout this process, varying formats are used by different research groups. Manual processing is normally involved in each step as data is moved from tool to tool. Key to streamlining the process, which would be very desirable, is to standardize the data.

Important areas for standardization include format, dimensionality, and metadata. Many formats exist for recording the same type of data, and is usually decided by the software used for processing or recording it, so no one format is always used. In most cases it may also be necessary to record the dimensionality of data; this includes the time and place data was taken – i.e., the metadata used to describe these attributes. This is useful for finding the right data, however, confusion arises when sharing data between different groups due to the varying formats and metadata structures during retrieval. Therefore, standardizing both the data format as well as the metadata are critical for sharing data among tools. Metadata representing the provenance of the data would represent how the data was collected, pre-processed, or reformatted.

Another important issue is data discovery and retrieval. Data can be mined using web services such as ftp or OPeNDAP [Ornillon et al 2003]. OPeNDAP is used for scientific data networking and ideal for environmental data. Data can be easily stored, even in different formats, however, retrieval becomes an issue when a standard is not put in place. How should data be stored (locally, distributed, centralized)?

Standardization remains a challenge. Several factors need to be considered when choosing standard data formats and metadata structure, such as: Should we only use open source formats? Should we use text formats, which are easier to understand but slower to process? Should we have one format for each data type, or choose a couple of widely accepted formats? Should we base our format choices on most common software, such as ArcGIS? What information is necessary in the meta-data?

In conclusion, data standardization allows for better data querying and less confusion in data usage.



### 3.3 Publishing Data and Publishing Models

Data generated by the individual scientist are often considered intellectual property and are not shared or published, which constitutes a tremendous loss for the scientific community. Incentives for the individual researcher are needed to overcome this cultural problem [Borgman et al 2012].

In addition to facilitating the process of data sharing by providing easy means of standardizing, annotating, and storing data, there should also be extra benefits that make sharing data not a duty but an investment for the individual scientist. Ideally, data sharing tools should provide the data in different standardized formats, automatically carry out standard data analysis, and produce simple plots.

In addition to facilitating the process of data sharing by providing easy means of standardizing, annotating, and storing data, there should also be extra benefits that make sharing data not a duty but an investment for the individual scientist. Ideally, data sharing tools should provide the data in different standardized formats, automatically carry out standard data analysis, and produce simple plots. By employing standard tools the researcher can also be informed of similar data for comparison, and be advised of further analysis based on the type of data recognized.

Data sharing would be facilitated by:

1. Tools that allow for easy standardization, annotation, and sharing of data while providing extra benefits as an incentive, such as reducing the cost of doing data conversions, plots, and standard multi-step analyses.
2. A convention/standard for associating metadata and provenance with data files similar to the Exif standard for image files and the NetCDF standard for geophysical data files [Rew et al 2006] and the ISO 19115-2 Lineage section [ISO 2003].
3. Tools that establish provenance by automatically recording processes used for data calibration, cleaning, and analysis, as well as handling and storing them using those standards
4. Provenance plug-ins for existing tools that are widely used in the scientific community (e.g. Excel, Matlab, Python, R, etc.) or a new generation of provenance-aware tools with similar analytic capabilities and uptake.
5. A peer-review system for publishing data, which will be stored as citable units in a repository (possibly as a new category of “data articles” in a journal). As the scientific community will start to value such “data publications” towards the performance of individual scientists, this will provide a tremendous incentive for data sharing.

Software sharing is just as important as data sharing but seldom recognized as a desirable practice. While data products created with numerical simulations are generally treated similarly to measured data, the notion that numerical simulations or “models” themselves should be shared as well is only slowly starting to gain traction in the community. Model details, underlying assumptions and the associated scope of applicability are often either hidden behind proprietary or badly documented and unintelligible code. This hampers keeping track of the provenance of data products generated with the models and therefore their interpretation and reuse. It also renders publishing and disseminating the models themselves extremely difficult. This is especially problematic in the light of technology export to developing countries. Any software and codes used to process data could be reused by others, and their publication should be encouraged and rewarded.

Software sharing would be greatly facilitated by:

1. Tools and standards for the unambiguous, explicit, consistent, intelligible, and well documented and annotated (with underlying assumptions) definition and implementation of models (numerical simulations).
2. Standards, central repositories and/or other ways for publishing, storing and disseminating those models (just like data): as a means of conveying the provenance of model data products, and for providing a library of models that can be easily adapted for similar systems all around the world.
3. Tools, services, and standards to allow for or facilitate the migration of organically grown, decades old, large pieces of valuable model code in outdated languages to the above described new structures.
4. Tools that allow for a rigorous consistency check of all aspects of a numerical model that is to be shared by a large user base.

## 4 Common Themes and Relevant Technologies: New Opportunities

Workshop participants discussed the role of recent research in computer science in addressing the above challenges, including interactive data integration tools, workflow systems, open semantic metadata annotation, and provenance-aware systems.

### 4.1 Data Cleaning and Normalization

Ecologists spend significant time collecting data in the field, supplementing it with data in shared repositories and then preparing it to be useful for running computational models. Raw data from sensors needs to be cleaned to remove noise and spurious data points; sometimes sensor calibration drift necessitates systematic adjustment of the data. Shared datasets often need to undergo cleaning too given that data may have not been properly cleaned before uploading, and also because downloading and extraction may introduce artifacts that need to be removed. Once datasets are cleaned, they may need to be normalized so that, for example, all data sets use the same units for the same measurements. After normalization, the data sets can be integrated to produce the data sets in the formats required in the modeling software.

Interactive data editors such as Microsoft Excel and plain text editors are popular tools to perform the cleaning and normalization data preparation steps. These tools are popular because they are visual and easy to use. For example, Excel shows the data in a familiar table format, offers convenient tools to make both ad hoc and systematic changes to data, and users immediately see the effects of their modifications. The main drawback of these interactive tools is not that the process is tedious and time-consuming, especially for large data sets. The main drawbacks are that the process is not repeatable and that no provenance information is left behind. Repeatability is important because many data sets are collected periodically, and the cleaning and normalization operations need to be performed again and again. Repeatability ensures the process is

While data products created with numerical simulations are generally treated similarly to measured data, the notion that numerical simulations or “models” themselves should be shared as well is only slowly starting to gain traction in the community. Model details, underlying assumptions and the associated scope of applicability are often either hidden behind proprietary or badly documented and unintelligible code. This hampers keeping track of the provenance of data products generated with the models and therefore their interpretation and reuse.

performed identically in each period, and once systematized, it can be performed quickly. Data cleaning and normalization steps can introduce errors and biases. The interactive tools do not record provenance information during these operations, so it is difficult to trace data sets back to the original raw sources. Finally, without insight into these processes it is hard for others to understand and reuse published data.

A new breed of tools aim to provide the ease of use benefits of the interactive data editors and the repeatability, efficiency and provenance support of scripting tools. The idea is to provide a visual interface where users build their script one small step at a time, and where the effects of each small step are immediately visible, like in a spreadsheet.

One alternative to these interactive data editors are scripting tools such as R, Perl and Python. The advantage of these tools is that once developed, the scripts can be efficiently applied to large data sets and to new versions of data sets collected periodically. The drawback of these tools is that they require programming expertise to use effectively. Many users who face data preparation tasks are not trained in programming and are unable to use these scripting tools.

A new breed of tools such as Google Refine [Google Refine 2011], Data Wrangler [Kandel et al 2011], and Karma [Tuchinda et al 2011] aim to provide the ease of use benefits of the interactive data editors and the repeatability, efficiency and provenance support of scripting tools. The idea is to provide a visual interface where users build their script one small step at a time, and where the effects of each small step are immediately visible, like in a spreadsheet. Google Refine provides a metaphor similar to entering formulas in Excel, using a specialized language for building the script elements. Data Wrangler and Karma go one step further and enable users to provide examples of data transformations from which these tools infer general procedures. These tools enable users to chain multiple transformations, and produce scripts that can be stored and reused. For example, the user could demonstrate how the first entry in a column of should be normalized. The system creates a generalized procedure to normalize dates, and applies it to the rest of the entries in the column. If the effects are not what the user wants, he or she can provide another example, and the system can use all the examples provided to infer an appropriate procedure. Data preparation steps can be supported using this by-example metaphor, resulting in tools that is both easy to use like MS Excel, but offer the power, repeatability and provenance benefits of scripts.

## 4.2 Workflows

Scientific workflows can be developed to automate data flow and analysis and manage metadata and provenance, thereby helping scientists to accelerate through these time- and human resource-intensive aspects of the scientific process. Sophisticated examples of workflows have existed in the ocean sciences community for some time (e.g., [Howe *et al.* 2008; Ramp *et al.* 2009] and have begun to emerge in support of ecological and hydrologic sensor network efforts [Barseghian *et al.* 2010; GLEON CDI 2011, Horsburgh et al. 2011], in some cases automating data analysis processes through scientific workflow systems [Gil et al 2011a].

Scientific workflow systems are becoming an enabler of complex scientific analyses (Taylor et al 2007). They provide a representation of complex analyses composed of heterogeneous models designed by several scientists. At the same time, workflows have also become a useful representation that is used to manage the execution of large-scale computations. This representation not only facilitates overall creation and management of the computations but also builds a foundation upon which results can be validated and shared. Since workflows formally describe the sequence of computational and data management tasks, it is easy to trace back how particular data were derived (i.e., the provenance). Workflows have also become a tool capable of bringing sophisticated analysis to a broad range of users, enhancing scientific collaboration and education.



In order for facilitate workflow creation, scientists need to be allowed to formulate the workflows in a way that is meaningful to them using high-level abstractions that specify the overall structure of the analysis and the data to be operated on (via a visual or textual representation) in a resource-independent way. This abstract workflow is important because it uniquely identifies the analysis to be conducted at the application level without including operational details of the execution environment. The workflow can thus be published along with the results to describe how a particular data product was obtained. Some workflow systems such as Wings [Gil et al 2011b] allow the user to specify the workflow at a high-level of abstraction, relying on semantic technologies to represent domain concepts and constraints and to reason about them to validate, elaborate, and suggest workflows.

Workflows not only facilitate overall creation and management of computations but also provide a foundation upon which results can be validated and shared. Since workflows formally describe the sequence of computational and data management tasks, it is easy to automatically record provenance and therefore trace back how particular data were derived.

In order to support the abstract workflow specifications, which let scientists concentrate on the science rather than on the operational aspects of the cyberinfrastructure, mapping technologies are needed to automatically interpret and map the user-defined workflows onto the available resources. This is an approach analogous to traditional computer programming methods, where high-level languages are used to describe the computation without needing to specify the use of specific registers or memory locations. In this analogy, the “workflow mapping engine” is a compiler that translates between the high-level specifications and the underlying execution system and optimizing the executables based on the target architecture. The mapping includes finding the appropriate software and computational resources where the execution can take place as well as finding copies of the data indicated in the workflow instance. The mapping process can also involve workflow restructuring geared towards optimizing the overall workflow performance as well as workflow transformation geared towards data management and provenance information generation. In some cases, the mapping is part of the workflow design process and is conducted by the user (e.g., Kepler [Ludäscher et al 2006] and Taverna [Oinn et al 2006]). In other cases, the mapping onto resources is done automatically (e.g., Pegasus [Deelman et al 2005]).

The result of the mapping process is an executable workflow, which can be executed by a workflow engine that follows the dependencies defined in the workflow and executes the activities defined in the workflow nodes. For example DAGMan [Couvares et al 2007], the workflow engine behind Pegasus, relies on the resources (compute, storage and network) defined in the workflow to perform the necessary actions. As part of the execution, the data is generated along with its associated metadata and any provenance information that is collected.

The separation of concerns between workflow creation, workflow mapping, and workflow execution allows for the design software in a modular way and to optimize the components based on their functionality.

Workflows can be very complex, encompassing millions of computational tasks, or simple computational pipelines with just a couple elements (for example data reformatting and visualization). In either case, the ability to capture metadata and provenance information for these workflows is critical in being able to find and interpret their results.

### 4.3 Metadata and Open Data Publication on the Web

Although the benefits of metadata are well recognized, the management of metadata in day-to-day science is far from ideal. First, many important metadata are buried in notebooks, emails, documentation files, or even worse not recorded in any form. When metadata is recorded, it is often

done manually when a dataset is deposited in a repository. Data repository managers define a comprehensive metadata record form that contributors must fill out painstakingly, often with limited understanding of its benefits and use. One of the disincentives to recording metadata is that when data is retrieved from those repositories to be further analyzed, the new data products must be annotated manually again.

Most metadata recorded voluntarily and informally typically contains the author and the timestamp. Other metadata is known to a scientist implicitly, for example by knowing who collected the data they would know the location where it was collected.

Other types of metadata would also be necessary. Of particular interest is semantic metadata that describes the types and properties of data. For example, whether a dataset is temperature and where it is Celsius or Fahrenheit. Semantic metadata is tremendously facilitated by the RDF web standard.

The Linked Data paradigm presents new opportunities for data sharing and publication in science through open web publication and semantic metadata annotation. Any scientist could publish a dataset as a web resource, and create metadata in RDF, using their own terms and ontologies (reusing community ontologies when appropriate). Emerging metadata representations could arise from this new open publication framework.

RDF is already being adopted by some environmental research infrastructure. In RDF, knowledge is structured in the form of object-property-value triples that can be aggregated to answer structured queries. The properties are expressed in terms of an ontology, which can be either a community-developed one or one developed by the individual scientist. There are many tools for ontology mapping and integration, which facilitate the integration of data.

Traditionally, this kind of metadata has been imposed by the developers of the shared repositories. That is, a comprehensive

metadata schema is created, and data publishers are required to provide all the fields, which may be irrelevant to their data, cryptic, or tedious.

Recently, the RDF standard together with a few simple publication principles have given rise to a new data publication movement on the Web. The idea of Linked Data was proposed by Web inventor Tim Berners-Lee as a complement to the current web that links documents. He proposed key principles for publishing open data on the web (<http://www.w3.org/DesignIssues/LinkedData.html>), so that data is published as openly accessible web objects, using RDF standards, and linking to other data. RDF links enable navigation from a data item within one data source to related data items within other sources using a browser. RDF links can also be followed by the crawlers of Semantic Web search engines, which may provide sophisticated search and query capabilities over crawled data.

This has resulted in a data commons that contains more than 200 datasets with 25B of interrelated facts contributed voluntarily by diverse communities [Bizer et al 2009]. There are standard web tools to find data, to reason about their metadata properties, and to query them based on the ontologies stated within. Its contents include BBC programming, New York Times subject headings, CNET product data, publications, friend-of-a-friend networks, the PubMed bibliography and many more. Many of these datasets have relevance to the ecological research community. As an example, the website LinkedGeoData published a Linked Data version of the OpenStreetMap dataset adding around 2 billion triples. It also contains DBpedia [Auer et al 07], an automatically extracted dataset of facts from Wikipedia's infoboxes and that contains 769,000 triples. In May 2010, following the presidential directive on Open Government, the US government portal Data.gov made around 400 of its datasets available as Linked Data, summing up to 6.4 billion triples (The UK government has embarked in a similar effort). Many biomedical datasets are also available, including the Protein Data Bank and the These datasets are interlinked with one another. For example, the Thomson Reuters

Web service that automatically generates semantic metadata for content (Calais) now supports Linked Data for all identified entities and links to DBpedia, Geonames, and other data assets in the cloud. The web of Linked Data continues to grow not just in size but also in breadth and coverage.

The Linked Data paradigm presents new opportunities for data sharing and publication in science through open web publication and semantic metadata annotation. Any scientist could publish a dataset as a web resource, and create metadata in RDF, using their own terms and ontologies (of course reusing community ontologies when appropriate and worth the trouble). Emerging metadata representations could arise from this new open publication framework.

#### 4.4 Provenance-Aware Tools

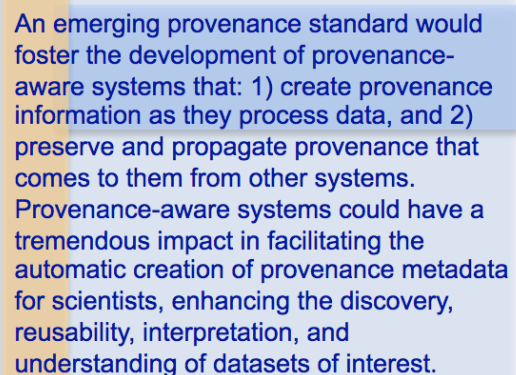
Provenance metadata is particularly crucial, as it refers to the record of all steps and processes that were applied to some initial data in order to obtain a result. Provenance metadata could be captured automatically by software, but most tools that scientists use do not capture provenance.

Provenance is most useful to those who are not intimately familiar with a dataset and therefore do not know implicitly what its provenance is. Unless provenance is represented uniformly across tools, it would be impossible to understand it and query it by third parties. One key challenge is that provenance is understood in very different ways across many areas of computer science, library sciences, and other disciplines. As a result, very diverse views and emerging standards for provenance have been proposed with varying degrees of adoption. Best known from library sciences are the Dublin Core Metadata [DC 2011] and the Premis Vocabulary [PREMIS 2008] by the Library of Congress, the former emphasizes source attribution and the latter version relationships. The scientific workflow community has developed jointly the Open Provenance Model (OPM) [Moreau et al 2011], focused on processes. Other provenance vocabularies have been developed in specific communities of practice.

A new opportunity in this area is the results of a community effort organized under the auspices of W3C, which collected use cases for provenance across different communities, surveyed the state of the art, analyzed immediate needs for standardization, and proposed 17 terms as the core of a new standardization effort [W3C Provenance 2010]. Since April 2011, a W3C Working Group is pursuing this standard effort [W3C Provenance 2011], and within a few months there will be common mechanisms to represent provenance. This will create new opportunities, as this standard will enable exporting, querying, and integrating provenance records across different tools.

An emerging provenance standard would foster the development of *provenance-aware systems* that: 1) create provenance information as they process data, and 2) preserve and propagate provenance that comes to them from other systems.

Provenance-aware systems could have a tremendous impact in facilitating the automatic creation of provenance metadata for scientists, enhancing the discovery, reusability, interpretation, and understanding of datasets of interest.



An emerging provenance standard would foster the development of provenance-aware systems that: 1) create provenance information as they process data, and 2) preserve and propagate provenance that comes to them from other systems. Provenance-aware systems could have a tremendous impact in facilitating the automatic creation of provenance metadata for scientists, enhancing the discovery, reusability, interpretation, and understanding of datasets of interest.

## 5 Observations and Insights

Some of the observations and identified challenges that emerged from the workshop discussions are new and seem to have unique prominence in environmental sciences research. Although some others may not be new, we include them here to highlight the importance of finding new approaches and new solutions.

### 5.1 Community Needs

Workshop participants agreed to several observations about the needs in the community:

- **Scientists in environmental sciences have varying degrees of resources**, many of them have very limited resources (e.g., one person shops) and yet by being very focused on a locality they develop a treasure trove of data and insight on a particular data point of overall ecosystems. The cost of sharing these datasets makes it impractical for them to publish the data in ECO repositories. This is a unique feature of environmental sciences that must be recognized, and appropriate opportunities for these scientists to participate in overall community efforts must be facilitated.
- **Many environmental scientists are unaware of relevant advances in computer science** that would be very appropriate for their needs, because it is hard to have visibility into rapidly changing areas of computer science that are not traditionally connected with environmental sciences. This includes research in data and information management systems, intelligent assistance, and collaborative problem solving infrastructure.
- **Challenges and technology requirements are shared across research areas in environmental sciences**. There are not many forums for the community to come together and articulate and share their needs. When a forum such as this workshop is facilitated, the commonalities are palpable.
- **There are many datasets collected and owned by individual scientists in a variety of areas in environmental sciences that are extremely valuable and yet are not shared**. Given the investment to date in large cyberinfrastructure efforts, there is a question of the cost that these scientists would have to export their data given their limited resources and practical lack of incentives. Reducing the cost of publishing datasets would open the doors to a torrent of data that is invaluable for environmental science research but is currently bottled up.
- **Scientists collect datasets separately, and invest enormous amounts of time organizing and preparing data that have in practice slightly altered formats and are quite similar in nature**. Spreadsheets are widely used because the user interface is very accessible and they are present in virtually every personal computer, but are far from ideal for the tasks that scientists face. Better frameworks for these data preparation and integration activities are needed.
- **The difficulty of finding datasets is so high that many opportunities are lost, even though there are many datasets that do exist out there that would benefit researchers**. Reusing data is also hard, in many cases the metadata is not useful and does not help in interpreting and integrating a dataset.
- **Workflows have been defined in several communities and are used for explicit process sharing with a number of benefits**. Scientists find they facilitate the organization of community efforts, since they make explicit what work needs to be done and how everyone's contributions fit in the context of the overall workflows. Finding ways to manage these workflows in more automated and assisted ways through workflow systems would be very beneficial in streamlining collaborative efforts in the community.
- **Environmental science would benefit for many reasons from faster turnaround from sensor to analysis**. It is easy to collect large volumes of data in a field campaign, but analysis can take many weeks. These problems often amount to creative data wrangling. Data wrangling may

transform data into a form that a particular scientist can use; however, even after the effort to transform the data, other scientists cannot leverage that work. Often, metadata is lost in the transformation.

- **The continuity of provenance and other metadata** would not only improve the usefulness of the data to the initial scientist, but also add value in enabling the reuse of the transformed data by other scientists.
- **Remote sensing is positioned to have an immediate impact in environmental sciences if its use were better supported by infrastructure.** It is often hard to locate, interpret, and integrate with local datasets.
- **Facilitating the creation of shared data formats and metadata properties would be very beneficial.** There is a clear cost of community investment and potential lack of agreements and delays in making progress. However, scientist face many data integration and interpretation tasks that would be greatly facilitated from shared representations. Approaches to facilitate these social processes would be very beneficial.

## 5.2 Perceived Technology Needs

Individual scientists often spend too much time solving IT issues that can dramatically cut in to the amount time they have for exploratory science and discovery. Scientists spend inordinate amounts of time on maintenance and development of ad hoc codes, gathering and preparing data, understanding the data (units, error characteristics, spatial/temporal characteristics, uncertainty), developing knowledge of multiple data formats and metadata standards, utilizing models, and running analysis codes many times over, often tweaking code on a case-by-case basis. This process can be inefficient and time consuming, can exponentially increase the time it takes to go from hypothesis to publication, is not conducive to maintaining provenance, does not support reuse of code, and can inhibit reproducibility of results.

Cyberinfrastructure is seen as a major part of the solution to the above problem, but scientists are typically viewed as consumers of it and often lack the resources and have little motivation to use shared data and tools. In an ideal world, scientists should be active participants and contributors to cyberinfrastructure and cyberobservatories to ensure that the broader environmental research community utilizes data gathered from individual measurements and in situ sensors. However, the reward/merit system for scientists is directly related to their publication record, usually within a narrow range of domain specific publications that do not provide an avenue to reward scientist participation in ECO projects.

Several important technology needs were identified:

- **Scientific workflows have the potential to improve the exploratory scientific process by providing a framework to organize and formalize standard methods for preparing data and performing analysis.** Through the use of workflows, metadata can be automatically generated and processing steps recorded ensuring provenance and providing the potential to breakdown data silos, preventing loss and underutilization of existing data. A scientific workflow framework can be used to break up existing processing steps in a way that supports modularity of codes (“plug and play”) allowing scientists to utilize different processing steps and components as needed, it encourages re-use of analysis (“solve once”) and data, and can help to ensure reproducibility of results.
- **Metadata and provenance capture should be done automatically as much as possible, because in practice the manual collection of metadata is not a practical approach.** There are many benefits to having metadata describing how data is collected (e.g., location, sensor characteristics), processes followed (e.g., normalization or cleaning steps), and other general properties of datasets. The benefits are clear, but scientists often face costly manual processes filling out lengthy forms that are hard to understand and that have unclear value added to them.



Metadata and provenance should be captured in every tool used in scientific data analysis processes.

- **Data repositories could be improved to push data to scientists, facilitating finding and reuse of datasets.** Although many repositories are available, they are organized for people to search, find, and use the data themselves manually. Better mechanisms for finding and reusing data in a way that is better supported by tools would be very beneficial.
- **Sharing and reuse of software for carrying out data analysis and simulations needs to be better encouraged and supported** with infrastructure. There are many models and tools that could be reused by different groups, and that could be used as steps to assemble workflows that do more comprehensive tasks. Similar issues to the sharing of data are applicable here: how to find, understand, reuse, integrate software from other scientists. While data repositories and data reuse have received significant attention in the past, the reuse and sharing of software has not.

**There is a clear theme concerning the need to support individual scientists in a more comprehensive and integrated way to carry out every task of the scientific analysis process: finding data, interpreting it, integrating it, processing it, and publishing it.** Many of these tasks remain manual, expensive, and not easily repeatable.

**Process-centered, comprehensive environments that support the flow of data and processes could pave the way to addressing bigger continental-scale interdisciplinary questions by facilitating data sharing across research groups and disciplines and removing the time consuming process of managing data analyses by hand.** This would enable distributed team science through its capacity to formalize methods, track provenance and metadata, and enable sharing of data and processing codes.

### 5.3 Collaboration Challenges

Experience tells us that there are many benefits but also many challenges in multi-disciplinary research involving computer science (CS) researchers with environmental sciences and other disciplines.

Publication outside of computer science venues is not counted for the purposes of career advancement. Multi-disciplinary collaborations often result in significant software development or system building focus with little basic innovations and therefore are not easily published in traditional computer science venues.

The difficulty of forming academic partnerships between computer science and other scientific disciplines can be circumvented by allowing students to earn credit by working or going into the field with the scientist to support data management activities. If the work is being done at a laboratory rather than at a university, interns or post-docs are ideal candidates for this sort of work. By pairing cyberinfrastructure savvy computer science students with ecologists, data acquisition and transformation systems can be quickly developed. With those systems as a basis, computer science researchers can then be involved in extending them in ways that require novel research.

The timescales of the research of environmental scientists and computers scientists are often at odds. Computer scientists work at a faster pace, from problem to solution to publication in 6 months to a year. Environmental scientists invest a lot of time in collecting data, and then in analyzing it, producing publications in the 2-year out timeframe. This often makes the collaborations loose momentum or be hard to synchronize to be productive in the two different disciplines.

Some suggestions for moving forward include building capacity for the future, encouraging interdisciplinary projects, and development of a community of practice.

1. **Building capacity today for the future:** Encourage/foster collaborations at the university level, for example with courses and credits that pair CS students w/science students interesting problems. It is important to get CS students into the field.
2. **Encouraging inter-disciplinary projects:** Inter-disciplinary (CS and environmental science) projects are key for moving forward and garnering community support. Identify opportunities for inter-disciplinary projects (need to identify “early adopter” scientists). Rewards need to be in place for cross-disciplinary research.
3. **Development of communities of practice:** Create a community of interest to connect those who are interested in adopting new technologies with those who are taking it upon themselves to become experts in their usage. The intention is to start a conversation about how these technologies can support research scientists in a more expeditious manner. The community of interest, could provide a mechanism for articulating uses of these technologies and help to identify collaborators, develop partnerships, and identify funding opportunities.

## 6 Vision: Enabling New Aquatic Ecosystem Science

Workshop participants discussed forward-looking scenarios for aquatic ecosystem research. We present here a vision for ecosystem science and for new cyberinfrastructure.

### 6.1 Enabling New Aquatic Ecosystem Science: Nitrogen and Carbon Dynamics

As a working example, we pose a visionary scenario for an aquatic ecosystem science question sufficient in both scale and complexity to require acquisition, analysis and synthesis of a broad spectrum of legacy and modern CO data. The science question is typically posed as a hypothesis that, for complex environmental systems, is often summarized as a conceptual model that encapsulates the underlying processes. The workflows for addressing this science question involve automating the parameterization, calibration, and testing of this model-encapsulated hypothesis. In this example, the workflow was directed at investigating current nitrogen and carbon dynamics in an estuarine ecosystem together with the associated management-relevant quantity of the estuarine filter function, i.e., the amount of carbon and nitrogen that is exported to the atmosphere within the estuary (akin to the carbon and nitrogen buffering capacity of the estuary). This workflow might also be framed so as to automate testing of a specific hypothesis related to nitrogen and carbon dynamics, for example: *Hypothesis: Estuarine algal blooms are controlled by the hydrograph and water quality of the river (which are coupled to land use in the watershed).* The resulting overall workflow sketched by the workshop group is summarized in Figure 1.

For the sake of brevity, we do not consider here details regarding the granularity of coverage, such as the number and spacing of instrumented stations. Instead we focused on the types of data expected. A first step toward solving the task at hand would be to obtain and collect previous knowledge and specific data from the system in question. In a multi-scale sensing context, this would include:

- (1) **Historical/Legacy Data.** These include data defining the geometry (or other inputs) of the estuary and river channel, including geomorphological data such as bathymetry and sediment types, water quality data derived from past sampling approaches and other information which will help to inform the design of the observational network. Remote sensing or related geospatial products, such as the National Land Cover Database (NLCD) may also be useful (e.g., relating land use changes to estuarine processes). Much of this data is available in the scientific literature or government reports and websites. An obvious issue with many of the historical data sets is that they are usually measured on an irregular basis (e.g., NLCD is available for 1992, 2001 and 2006). Digitizing, standardizing and updating these heterogeneous datasets requires substantial effort. When it is done once, the results should be easily shared.

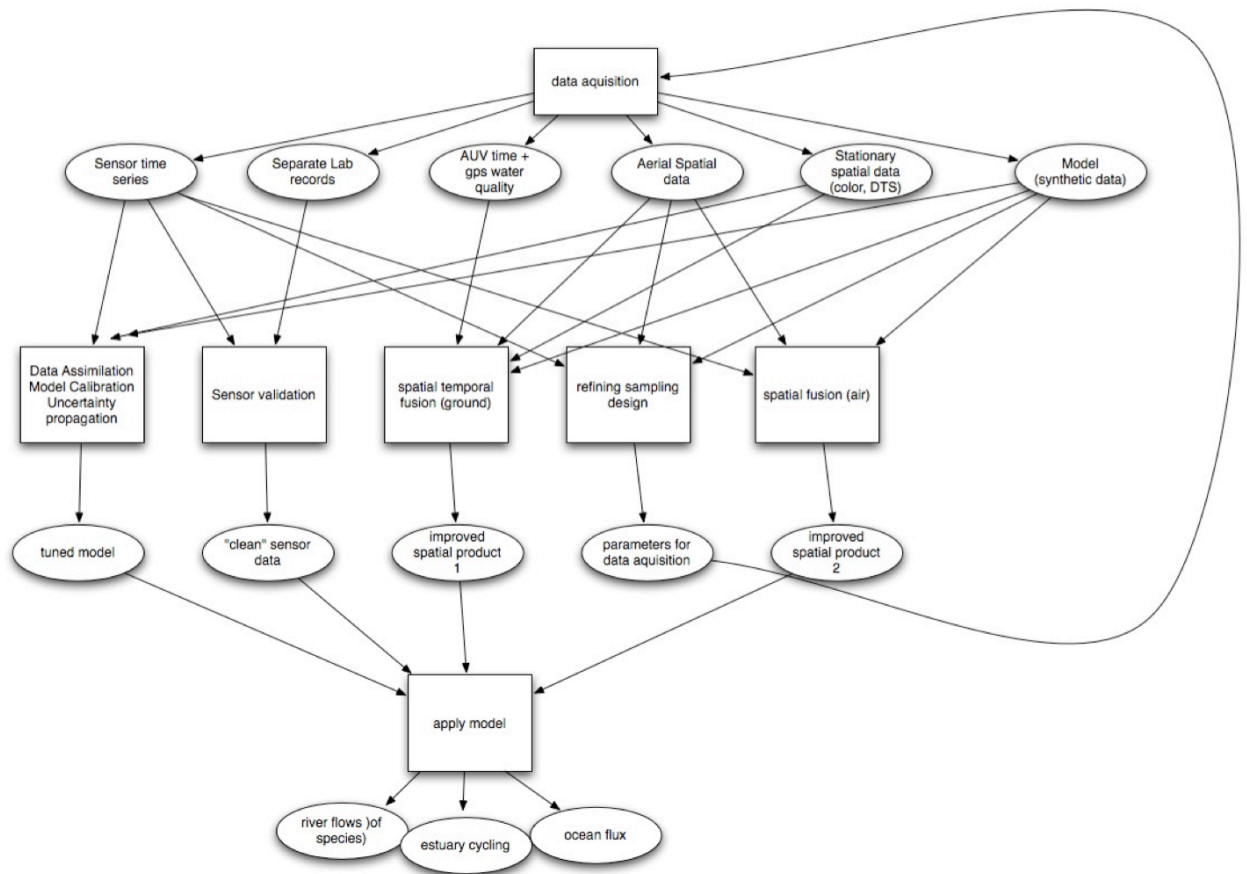


Figure 1: A sketch for a workflow for nitrogen and carbon dynamics in an estuary ecosystem.

- (2) **Traditionally Sampled Data.** These data include water grab samples analyzed for parameters such as trace nutrients, metals, organic microbial species and enumeration. Again, digitizing, standardizing, updating, and integrating these data requires substantial effort.
- (3) **Stationary *In-Situ* Sensor Data.** These data are from instrumented buoys or river gauging stations equipped with sensors monitoring local water conditions (temperature, salinity, dissolved oxygen, and sometimes other parameters, such as pH, nitrate, total suspended solids (TSS), dissolved organic matter (DOM), chlorophyll-a, and point velocity (ADV) or velocity profiles (ADP). Often these stations also monitor meteorological conditions (air temperature, precipitation, relative humidity, solar radiation, wind speed and direction). Similar data from nearby agency-maintained stations may be available. River and estuary gauging stations are also typically equipped with depth sensors which can be used to monitor tide changes or from which river discharge is estimated based on a periodically updated rating curve. These data are typically available as times series for a fixed location (the ADP data being an exception). Much of this data will be available online from government administered databases, and some of it streaming continuously. Obtaining and cleaning all necessary data, as well as incorporating them in a standardized format is still a largely manual, cumbersome process and represents a major bottleneck for the scientist.
- (4) **Mobile *In-Situ* Sensor Data.** These sensors are similar to those mentioned above but they are conveyed by human-piloted ships or AUVs over planned or adaptively managed



courses to provide a three-dimensional trace of the water conditions in the estuary.

- (5) **Remote Sensing (Aerial/Satellite and Land-Based) Data:** Satellite products from LandSat or MODIS are available at regular time intervals and are useful for defining land use and boundaries on a relatively coarse spatial scale. MODIS can also be used to determine ocean surface temperature and color related to suspended solids (TSS) and phytoplankton (chlorophyll). Aircraft-based sensors such as AVIRIS can be used to provide hyperspectral reflectance imagery at higher spatial resolutions, but significantly less frequently as a satellite, as these sensors are typically operated as shared systems. Aircraft-based LIDAR (including green LIDAR) is also becoming increasingly available and might be useful in the estuary scenario to delineate the micro-topography of the intertidal zones and perhaps the bathymetry of shallow portions of the estuary. Ground-based systems in this category include coastal radar (CODAR) stations which provide real-time ocean surface velocity fields, and possibly distributed temperature sensing (DTS) systems for delineating dynamic and spatially distributed phenomena like groundwater discharges into the estuary or intertidal boundaries. In general, this category of data includes large spatial data sets, such as images, radiometric spectra, and vector fields (e.g., CODAR surface velocities). Some of these data products are relatively standardized (e.g., satellite products) while others require significant effort including pre-processing (geo-referencing, radiometric correction) and post-processing (e.g., alteration detection, segmentation, classification).

Using the above data, domain scientists can formulate and test their hypothesis by implementing appropriate numerical simulations of the estuary. Simulations generate synthetic data describing the continuum of physical and biogeochemical properties and processes occurring in the estuary. Other synthetic data may be created using simpler statistical models, for example, to resample or interpolate between points in time series or synoptic data sets. With data acquired, cleaned, standardized, and used to develop the conceptual model of the estuary, the hypothesis can be quantitatively tested. Model parameters as well as model structures are iteratively refined until the model output data sufficiently matches measured monitoring data, meaning the hypothesis represented by the model is

In a broader scientific context, the impact of scientific workflows of the type envisioned here is difficult to assess, but likely to be transformative. These workflows will provide high quality, standardized data, integrated from heterogeneous observational perspectives as input for complex and flexibly structured environmental simulations. Many in the community use informal sketches of workflows already to describe these processes.

accepted to represent reality within an acceptable level of confidence or uncertainty. At this point the result of the scientific task at hand can be extracted, e.g., either in form of a confirmed or falsified hypothesis or in the form of ecosystem management relevant integrated data products. For the presented example, these results could be, e.g., estimates for the system-wide carbon and nitrogen cycling as well as the efficiency of the estuarine filter function. Furthermore the uncertainties associated with these estimates can also be determined.

Modeled synthetic data, however, may also be treated as an additional data source in the above list and can be employed to complete additional scientific tasks. With model data, challenges center around capturing and managing the

metadata needed to create these models, including their boundary conditions and spatiotemporally distributed parameters, and the provenance of the simulations, which range from preliminary tests, to calibration/assimilation, to hypothesis-testing or forecasting.

For assessing ecosystem metabolism and element cycling, the core process of 1) implementation of a hypothesis of ecosystem functioning via building a numerical model from known building blocks

based on domain expertise, and 2) iterative testing and refinement of the hypothesis by comparison of model output to measured, is the same for virtually any aquatic ecosystem. Furthermore, not only are similar types of data used, but also the toolbox of building blocks used to construct a numerical simulation is virtually the same for many different systems. Hence, what is needed is a commonality of building blocks for “plugging in” new observational components, including new types of measurements and expansion of existing measurements in time and/or space.

The whole aquatic biogeochemical modeling community would greatly benefit from a facilitation of workflow composed of tools that 1) provide high quality, standardized input data for numerical simulations, and 2) allow for the modular construction of numerical simulations via the selection of pre-coded building blocks, yet retain enough plasticity that new building blocks can be easily defined and added. In the context of the estuary hypothesis described above, it is likely that both a larger number and broader array of scientists (beyond the modeling community) would contribute to the developing and interpreting the resulting observations if sophisticated workflows consistently directed the hypothesis testing. Clearly, more eyes (and perspectives) on the testing and outcomes would result in a more rapid understanding of the system and therefore how to best manage resources connected to the system.

In a broader scientific context, the impact of scientific workflows of the type envisioned here is difficult to assess, but likely to be transformative. These workflows will provide high quality, standardized data, integrated from heterogeneous observational perspectives as input for complex and flexibly structured environmental simulations. Many in the community use informal sketches of workflows already to describe these processes. We have caught glimpses of how transformative the automation of such workflows may be when we view short-term weather reports. However, the biogeochemical processes at stake are much more complex than the weather, which draws largely from physical processes. With this in mind, it may be better to think of these workflows together with the data and metadata they transform as virtual multi-scale sensors, analogous to the first microscope. Few realized at the time of its discovery that the microscope would open an entirely new field of science. Just as few can now predict the impact of the “long tail” of science collaboratively engaged in cyberobservatory enabled inquiries.

## 6.2 Enabling Technologies: Envisioning a “Data Librarian”

We describe here a novel concept of a “Data Librarian” that illustrates a range of novel technologies that could be useful to the long tail of scientists. This system would interact with an individual scientist, and is more proactive than existing systems in providing assistance to publish, share, reuse, and analyze datasets. By doing so, the system is giving added value to the scientist in terms of an immediate reward that helps their work. The new capabilities are inspired by recent developments in machine learning and semantic technologies, and point to the possibility that a new generation of data management systems can be developed. We focus here on two key aspects of a Data Librarian: 1) the ability to assist in the creation of metadata, 2) the ability to proactively suggest to the scientist items of interest to process the data. The former would lower the cost of metadata creation significantly, the latter would provide immediate benefits to the scientist for any investment they make in metadata creation.

When a scientist has a dataset and uploads it to a Data Librarian, the system would automatically propose semantic descriptions of the data and suggests adding them as metadata. For example, the system would detect, for example, that a dataset looked like weather station data, and automatically identify some of the columns as temperature and humidity. This kind of capability could be developed based on machine learning techniques developed in recent years [Carman and Knoblock 2006; Dereszynski and Dietterich 2011]. The basic idea is that if the system has seen data of a certain type before, and has been told what type it is, it could use that information to automatically build recognizers to assign that type to new data.

A Data Librarian would help with quality control by detecting values that are out of range and suggesting common fixes. For example, a temperature reading may be impossible, a Data Librarian would flag it to the user. The user could fix it by entering the average value of the prior and subsequent readings, and indicate that this fix should be used to handle other incorrect values. A Data Librarian would automatically generalize the fix and apply to other values out of range. It would also add the fix to a library of common quality control techniques that it can offer other users as they clean their datasets.

The novel concept of a **Data Librarian** that illustrates a range of novel technologies that could be useful to the long tail of scientists. We focus here on two key aspects of a Data Librarian: 1) the ability to assist in the creation of metadata and provenance, 2) the ability to proactively suggest to the scientist items of interest (such as other relevant data as well as workflows) to process the data. The former would lower the cost of metadata creation significantly, the latter would provide immediate benefits to the scientist for any investment they make in metadata creation.

A Data Librarian would also automatically assign metadata to future datasets that the scientist uploads based on context and history of prior uploads by the scientist. For example, suppose a scientist uploads a daily dataset and creates metadata specifying the sensor and the site. The system would assume the same sensor and site for future data uploads. If a scientist uploads several datasets and adds metadata indicating that they have been normalized, the system could simply confirm with the user that future datasets were prepared in the same way and automatically add that metadata as well. A Data Librarian would apply all the quality control processes that the user has preferred in the past.

When a scientist takes the trouble to specify metadata about a dataset, the Data Librarian could suggest to the scientist relevant workflows that could be used to process their data. For example, a scientist uploading sensor data about water quality could be presented with workflows that the system knows about for doing metabolism calculations, and offer to run them automatically. This would give the scientist an immediate benefit for her effort in creating metadata: just by specifying the type of data that they have the system can help them run interesting analyses.

Throughout these processes, a Data Librarian would keep detailed provenance records of all the processes applied to the original data. Any manipulations of the data would be instrumented to allow the collection of these provenance records. This provenance would then be available so the system can retrieve it and suggest it for reuse and so other scientists can interpret the data. A Data Librarian would export data to other systems together with its corresponding metadata and provenance.

A Data Librarian would also use metadata to find other relevant data that would supplement the scientist's original dataset. For example, suppose a scientist uploads a dataset and describes it as water quality data collected from a specific station in a river. The Data Librarian would show them the workflow for calculating water metabolism together with a complementary dataset that it has access to in an ECO site has riverbed characteristics and that is needed to run the workflow in conjunction with the scientist's dataset. The Data Librarian might also suggest datasets from other users that contain similar sensor measurements and are in geographical proximity to the dataset uploaded by the scientist. For example, the system would show datasets from nearby locations in the same river, and suggest workflows with flow models that exploit data from several stations.

To facilitate stewardship of data and software, different instances and even different implementations of the Data Librarian software could be set up for different spheres of expertise and scope in the ecological research community. Data Librarians could communicate with one another to be aware of other community datasets, metadata and ontologies, and models and workflows.

## 7 Conclusions

This report discussed significant challenges in environmental sciences for the long tail of scientists that currently draw limited or no benefits from ECO investments. These include the effort required to find relevant ECO data, the difficulties in integrating their data with ECO data, and the cost of carrying out data analysis processes mostly manually. The report also proposed technologies to address some of these challenges through automation and assistance for semantic metadata annotations, workflow management, provenance recording, and semantic web publication technologies. The report also discussed observations and insights contributed by workshop participants on the community and technology needs, and the challenges for meshing those needs in a practical context. The report outlined a vision for future environmental science research practices, showing an example scenario for nitrogen and carbon dynamics and proposing a “Data Librarian” to illustrate the technical vision that would support the science scenario.

The key recommendations from the workshop participants are:

- **Data sharing approaches that reduce publication cost and provide immediate benefits to the scientist are important in order to rescue a lot of data in environmental science that will otherwise be lost.** Vast amounts of data are confined to the local file system of investigators that collect and curate readings often over many years. These datasets are not shared because of the cost involved in their publication. The development of systems that automatically add metadata through learning technologies would lower that cost significantly. In addition, new technologies can be developed to give scientists immediate rewards for their data publication efforts, by offering to do common quality control and data analysis processes on their data.
- **Workflow systems that manage and automate data processing steps are crucial to enable efficient processing of the volumes of data required to address continental-scale global-scale environmental science questions.** A lot of effort by scientists is misplaced on developing scripts, installing software packages, writing data cleaning and reformatting codes, and managing the overall orchestration of these individual steps. Instead, these processes should be shared in community repositories of workflows and software components, much like there are community repositories of data. In addition, creating and executing these processes could be managed through workflow systems, saving scientists time and recording automatically valuable provenance information.
- **More efficient processing of environmental data would improve data collection by enabling rapid adjustment of sampling and sensor configurations.** If data analysis processes are carried out in the order of days rather than months, then the scientists could adjust sensors in time to collect improved information about particular phenomena detected with their analyses.
- **Pervasive provenance recording would improve reuse and productivity by facilitating the flow of data and processes across research groups and disciplines.** Provenance records provide a context for any dataset that enables its reuse, as they describe how the data was collected, analyzed, and published. This can have a profound effect on facilitating the sharing of data across scientific disciplines.
- **Workflow sharing can drive and facilitate collaborative science projects that represent higher-impact science efforts.** Workflows express how datasets contributed by different groups are processed at different stages of an analysis. Workflow also express what steps are included in the process and clarify which groups are responsible for steps. Workflows effectively represent temporal dependencies that can help collaborative science projects. They also make explicit the role of the contributions by every participant.

- **Data and software repositories should exhibit an economy of scale where individual efforts save effort to others.** The investments of individual scientists should be leveraged so that others do not have to repeat work or incur costs that could be saved through reuse.

Although the backgrounds of the participants and the proposed discussion topics focused on aquatic ecosystems, the findings and recommendations of the workshop have broader applicability to ecological sciences and to science practices in all disciplines.

## 8 References

- Anderson, C. (2004). The Long Tail. *WIRED Magazine*, October 2004.
- Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. *Proceedings of the 6th International Semantic Web Conference (ISWC)*, Busan, Korea, (2007).
- Barseghian D., 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 (2010). Workflows and extensions to the Kepler scientific workflow system to support environmental sensor data access and analysis. *Ecological Informatics*. 5:42 - 50.
- Beasley, T. (2010). NEON data policy. National Ecological Observatory Network, Boulder, CO. Available: <http://www.neoninc.org/documents/463> [accessed June 2011]
- Bizer, C.; Heath, T.; and Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3), 2009.
- Borgman, C. (2012). Developing Data Attribution and Citation Practices and Standards, *National Academies Press*, to appear in 2012. Available from [http://sites.nationalacademies.org/PGA/brdi/PGA\\_063656](http://sites.nationalacademies.org/PGA/brdi/PGA_063656).
- Brown, L. R., T. F. Cuffney, J. F. Coles, F. Fitzpatrick, G. McMahon, J. Steuer, A. H. Bell, and J. T. May. (2009). Urban streams across the USA: lessons learned from studies in 9 metropolitan areas. *Journal of the North American Benthological Society*, 28:1051-1069.
- Buisson, L., W. Thuiller, S. Lek, P. Lim, and G. Grenouillet (2008). Climate change hastens the turnover of stream fish assemblages, *Global Change Biol.*, 14, 2232-2248.
- Capel, P. D., K. A. McCarthy, and J. E. Barbash (2008). National, holistic, watershed-scale approach to understand the sources, transport, and fate of agricultural chemicals, *J. Environ. Qual.*, 37, 983-993.
- Carman, M. J., and C. A. Knoblock (2007). Learning Semantic Definitions of Online Information Sources. *Journal of Artificial Intelligence Research (JAIR)*, 30:1--50.
- CZO (2010). Future Directions for Critical Zone Observatory (CZO) Science, Report prepared by the Critical Zone Observatories community for NRC NROES committee, December 29, 2010. Online at <http://criticalzone.org/>.
- Cornillon, P., J. Gallagher, and T. Sgouros (2003). OPeNDAP: Accessing data in a distributed, heterogeneous environment. *Data Science Journal*, Vol. 2, No. 0. (2003), pp. 164-174.
- Couvares, P., T. Kosar, A. Roy, J. Weber and K. Wenger, (2007). Workflow in Condor, *Workflows for e-Science*, I. Taylor, E. Deelman, D. Gannon, M. Shields (Eds), Springer, January 2007.
- DC (2011). Dublin Core Metadata Initiative. Available: <http://dublincore.org/>.
- de La Beaujardiere, J. (2008). The NOAA IOOS Data Integration Framework: Initial implementation report, *OCEANS 2008*, vol., no., pp.1-8, 15-18 Sept. 2008 doi: 10.1109/OCEANS.2008.5152007 URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5152007&isnumber=5151799>



- Deelman, E., G. Singh, M. Su, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, G. B. Berriman, J. Good, A. Laity, J. C. Jacob, D. S. Katz. (2005). Pegasus: a Framework for Mapping Complex Scientific Workflows onto Distributed Systems. *Scientific Programming Journal*, 13(3), 219-237, 2005,
- Dereszynski, E. W. and T. G. Dietterich. (2011) Spatiotemporal Models for Data-Anomaly Detection in Dynamic Environmental Monitoring Campaigns. *ACM Transactions on Sensor Networks*, 8(1), 2011.
- Dibike, Y. B. and P. Coulibaly. (2005). Hydrologic impact of climate change in the Saguenay watershed: comparison of downscaling methods and hydrologic models. *Journal of Hydrology*, 307, 145-163.
- EML (2011). Ecological Metadata Language. Available at <http://knb.ecoinformatics.org/software/eml/>.
- Gil, Y.; Ratnakar, V.; Kim, J.; Gonzalez-Calero, P. A.; Groth, P.; Moody, J.; and Deelman, E. (2011a). Wings: Intelligent Workflow-Based Design of Computational Experiments. *IEEE Intelligent Systems*, 26(1).
- Gil, Y.; Szekely, P.; Villamizar, S.; Harmon, T.; Ratnakar, V.; Gupta, S.; Muslea, M.; Silva, F.; and Knoblock, C. (2011a). Mind Your Metadata: Exploiting Semantics for Configuration, Adaptation, and Provenance in Scientific Workflows. *Proceedings of the Tenth International Semantic Web Conference (ISWC)*, Bonn, Germany.
- GLEON CDI (2011). GLEON Cyber-infrastructure Digital Innovation website: <http://cdi.gleon.org/>. [Accessed June 2011]
- Graham, M. J., Fitzpatrick, M. J., and T. A. McGlynn. (2008). The National Virtual Observatory: Tools and Techniques for Astronomical Research. *Astronomical Society of the Pacific (ASP) Conference Series*, Volume CS-382, 2008. Available from <http://www.astrosociety.org/CS382.html>.
- Hanson, P. C. (2007). A grassroots approach to sensor and science networks. *Frontiers in Ecology and the Environment*, 5(7), 343.
- Hanson, P.C. (2008). New ecological insights through the Global Lake Ecological Observatory Network (GLEON). *Ecological Science*. 27(5): 300-302.
- Horsburgh, J. S., D. G. Tarboton, D. R. Maidment and I. Zaslavsky, (2011). Components of an environmental observatory information system, *Computers & Geosciences*, 37(2): 207-218.
- Hodgkins, G. A., and R. W. Dudley. 2006. Changes in the timing of winter–spring streamflows in eastern North America, 1913–2002. *Geophysical Research Letters* 33:5 PP.
- Howe, B, Lawson P, Bellinger R, Anderson EW, Santos E, Freire J, Scheidegger C, Baptista A, Silva CT. (2008). End-to-End eScience: Integrating Workflow, Query, Visualization, and Provenance at an Ocean Observatory. *Proceedings of the 2008 IEEE Fourth International Conference on eScience*, 2008.
- IOOS-DMAC (2010). U.S. Integrated Ocean Observing System: A Blueprint for Full Capability. [http://www.ioos.gov/library/us\\_ioos\\_blueprint\\_ver1.pdf](http://www.ioos.gov/library/us_ioos_blueprint_ver1.pdf)
- ISO (2003) Geographic information – metadata. International Organization for Standardization. ISO 19115:2003. <http://webstore.ansi.org/RecordDetail.aspx?sku=ISO%2019115:2003>.
- Keller, M., D. S. Schimel, W. W. Hargrove, and F. M. Hoffman. (2008). A continental strategy for the National Ecological Observatory Network. *Frontiers in Ecology and the Environment* 6:282-284.

- Keller, M., Schimel, D.S., Hargrove, W.W., and Hoffman, F.M. (2009). A continental strategy for the National Ecological Observatory Network. *Frontiers in Ecology and the Environment* 6: 282-284.
- Keller, M. 2010. NEON level 1-3 data products catalog. National Ecological Observatory Network, Boulder, CO. Available: <http://www.neoninc.org/documents/513> [accessed June 2011]
- Keller, M., Alves, A., Aulenbach, S., Johnson, B., Kampe, T., Kao, R., Kuester, M., Loescher, H., McKenzie, V., Powell, H., and Schimel, D., 2010. NEON scientific data products catalog. National Ecological Observatory Network, Boulder, CO. Available: <http://www.neoninc.org/documents/513> [accessed June 2010]
- Knowles, N. and D. R. Cayan (2002), Potential effects of global warming on the Sacramento/San Joaquin watershed and the San Francisco estuary, *Geophys. Res. Lett.*, 29, 1891.
- Ludäscher, B., I. Altintas, C. Berkley, D. Higgins, E. Jaeger-Frank, M. Jones, E. Lee, J. Tao, Y. Zhao. (2006). Scientific Workflow Management and the Kepler System. *Concurrency and Computation: Practice and Experience*, 18(10), 1039-1065, 2006.
- Maidment, D.R. (Ed.) (2008). CUAHSI Hydrologic Information System: Overview of Version 1.1. Consortium of Universities for the Advancement of Hydrologic Science Inc., Washington, D.C., p. 92. Available from: <http://his.cuahsi.org/documents/HISOverview.pdf>
- Marcogliese, D. J. (2001), Implications of climate change for parasitism of animals in the aquatic environment, *Canadian Journal of Zoology-Revue Canadienne De Zoologie*, 79, 1331-1352.
- Monk, W. A., D. L. Peters, R. Allen Curry, and D. J. Baird. In press. Quantifying trends in indicator hydroecological variables for regime-based groups of Canadian rivers. *Hydrological Processes*.
- Moreau, L., B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. Stephan, and J. V. denBussche. (2011). The Open Provenance Model Core Specification (v1.1). *Future Generation Computer Systems*, 27(6).
- Morrison, J., M. C. Quick, and M. G. G. Foreman (2002), Climate change in the Fraser River watershed: flow and temperature projections, *Journal of Hydrology*, 263, 230-244.
- NEON 2009. The National Ecological Observatory Network (NEON) scientific strategy: enabling continental scale ecological forecasting. National Ecological Observatory Network, Boulder, CO. Available: <http://www.neoninc.org/sites/default/files/NeonScienceStrategySept09.pdf>
- NetCDF (2011). Network Common data Form. Available at <http://www.unidata.ucar.edu/software/netcdf/>.
- Newman, B. D., B. P. Wilcox, S. R. Archer, D. D. Breshears, C. N. Dahm, C. J. Duffy, N. G. McDowell, F. M. Phillips, B. R. Scanlon, and E. R. Vivoni (2006), Ecohydrology of water-limited environments: A scientific vision, *Water Resour. Res.*, 42, W06302.
- OceanObs (2009). Conference Statement of the Conference on Ocean Information for Society: Sustaining the Benefits, Realizing the Potential, <http://www.oceanobs09.net/statement/>
- Olson, G. M., Zimmerman, A., and N. Bos. *Scientific Collaboration on the Internet*. MIT Press, 2008.
- Oinn, T., M. Greenwood, M. Addis, N. Alpdemir, J. Ferris, K. Glover, C. Goble, A. Goderis, D. Hull, D. Marvin, P. Li, P. Lord, M. Pocock, M. Senger, R. Stevens, A. Wipat, and C. Wroe. (2006). Taverna: Lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience*, vol. 18, iss. 10, pp. 1067–1100.
- OOI (2005). Ocean Observatory Initiative Science Plan: Revealing the Secrets of our Ocean Planet. ORION Executive Steering Committee, Washington, DC.

- OOI (2010). Ocean Observatory Initiative Final Network Design. Version 2-06. [available at: [http://www.oceanleadership.org/wp-content/uploads/2009/04/1101-00000\\_FND\\_OOI\\_2010-04-22\\_ver\\_2-06\\_public1.pdf](http://www.oceanleadership.org/wp-content/uploads/2009/04/1101-00000_FND_OOI_2010-04-22_ver_2-06_public1.pdf)].
- Poff, N. L., J. D. Allan, M. B. Bain, J. R. Karr, K. L. Prestegard, B. D. Richter, R. E. Sparks, and J. C. Stromberg. (1997). The natural flow regime. *BioScience* 47:769-784.
- Poff, N. L., B. P. Bledsoe, and C. O. Cuhaciyan. (2006). Hydrologic variation with land use across the contiguous United States: Geomorphic and ecological consequences for stream ecosystems. *Geomorphology* 79:264-285.
- Porter, J., E. Nagy, P.C. Hanson, T.K. Kratz, S. Collins, and P.A. Arzberger. (2009). New Eyes on the World: Advanced Sensors for Ecology. *Bioscience* (59): 385-397.
- Puri, D., R. Karthikeyan, and M. Babbar-Sebens (2009). Predicting the Fate and Transport of E. coli in Two Texas River Basins Using a Spatially Referenced Regression Model, *JAWRA Journal of the American Water Resources Association*, 45, 928-944.
- Rew, R., Hartnett, E. J., and Caron, J. (2006) NetCDF-4: software implementing an enhanced data model for the geosciences, in: 22nd International Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology, <http://www.unidata.ucar.edu/software/netcdf/papers/2006-ams.pdf>, 2006.
- Quinn, N. W. T. (2011). Adaptive implementation of information technology for real-time, basin-scale salinity management in the San Joaquin Basin, USA and Hunter River Basin, Australia, *Agric. Water Manage.*, 98, 930-940.
- PREMIS (2008). PREMIS Data Dictionary for Preservation Metadata. Library of Congress Report, 2008. Available from <http://www.loc.gov/standards/premis/v2/premis-dd-2-0.pdf>.
- Ramp, S.R., P. Lermusiaux, R.E. Davis, Y. Chao, D. Fratantoni, N.E. Leonard, J.D. Paduan, F. Chavez, I. Shulman, J. Marsden, W. Leslie, and Z. Li, (2009). The Autonomous Ocean Sensing Network (AOSN) predictive skill experiment in the Monterey Bay. *Deep-Sea Research II*.
- Regonda, S. K., B. Rajagopalan, M. Clark, and J. Pitlick. (2005). Seasonal cycle shifts in hydroclimatology over the Western United States. *Journal of Climate* 18:372-384.
- Smith, J. W. N., M. Bonell, J. Gibert, W. H. McDowell, E. A. Sudicky, J. V. Turner, and R. C. Harris (2008). Groundwater/surface water interactions, nutrient fluxes and ecological response in river corridors: Translating science into effective environmental management, *Hydrol. Process.*, 22, 151-157.
- Stewart, I. T., D. R. Cayan, and M. D. Dettinger. (2005). Changes toward earlier streamflow timing across western North America. *Journal of Climate* 18:1136-1155.
- USGS (United States Geological Survey). (2011). USGS Surface water information. Resdon, VA. Available: <http://water.usgs.gov/osw/> [accessed June 2011]
- Utz, R. M., K. N. Eshleman, and R. H. Hilderbrand. (2011). Variation in physicochemical responses to urbanization in streams between two Mid-Atlantic physiographic regions. *Ecological Applications* 21:402-415.
- W3C Provenance (2010). Final Report of the W3C Provenance Incubator Group. World Wide Web Consortium Report, December 2010. Available from <http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/>.
- W3C Provenance (2011). W3C Provenance Working Group. <http://www.w3.org/2011/prov/wiki>. Accessed October 2011.



- Warren, N., I. J. Allan, J. E. Carter, W. A. House, and A. Parker (2003). Pesticides and other micro-organic contaminants in freshwater sedimentary environments - a review, *Appl. Geochem.*, 18, 159-194.
- WATERS, (2009). Living in the Water Environment: the WATERS Network Science Plan. Online at <http://www.watersnet.org/>.
- Woodward, G., D. M. Perkins, and L. E. Brown (2010). Climate change and freshwater ecosystems: impacts across multiple levels of organization, *Philosophical Transactions of the Royal Society B-Biological Sciences*, 365, 2093-2106.



**Information Sciences Institute**  
University of Southern California  
4676 Admiralty Way  
Marina del Rey, CA 90292  
<http://www.isi.edu>