

## 2.11 Statistics and data practices (SDP)

Data, statistical models and inferential procedures permeate CENS deployments, from the four founding scientific application areas to the more recent urban sensing campaigns and “n-of-one” personal data collection exercises. Rather than attempt to synthesize that work here, we will instead focus on three classes of activities not well represented elsewhere: 1) General statistical models for embedded sensing, with specific applications to data quality and continuous sampling, 2) Significant CENS-designed and supported databases and repositories, and 3) Studies into the data lifecycle for embedded sensing systems. On their own, these three tell a story of data sharing and reuse, of application-driven system design, and of the multiple levels of analysis that take place simultaneously in an interdisciplinary center like CENS.

### **General statistical models: Applications to data quality and continuous sampling**

#### *Fault and anomaly detection*

This year, our fault detection work proceeded on two fronts. In the first we considered various model-based approaches that flag as anomalous measurements that differ significantly from expectations. We report on two in our project summaries: A multiplicative seasonal ARIMA time series model, and a (Gaussian) Bayesian model designed to quantify “surprise” in one or more sensor readings. The second broad direction for our fault detection work is non-parametric in nature, depending on a set of well-chosen data features that are combined in a flexible “signature” that describes normal operation of a sensor or a cluster of sensors. Signatures are also constructed for various fault types, and, as new data are collected, these are compared to identify likely anomalous measurements. Both the model- and signature-based methods share common themes and technical challenges: Both depend on the underlying network topology for setting realistic constraints on data sharing between system components; both rely on local updates that allow the system to learn new, non-faulty behavior; both require well-documented deployment data from several application areas for training and experimentation; and both begin with some characterization of the fault process. In connection with this last point, we are pleased to report that our taxonomy of fault types, “Sensor Network Data Fault Types” will appear in the 2009 Proceedings of the ACM Transactions on Sensor Networks.

#### *A new sampling paradigm*

In applications involving mobile sensors, we have often carried over many notions that are rooted in static sensing. As a result, we often employ mobile nodes in a way that mimics static data collection: We move a mobile sensor to a grid point where a static sensor might have been, let it dwell in that location for some period of time until a sample is taken, and then move the node to another point. By contrast, over the last year, we have developed sampling schemes for situations in which we can sample (near) continuous paths using a single mobile node. Our test problem involves modeling dynamic light patterns in a forest understory. In addition to adapting many field estimation schemes to work with highly structured input data, we are developing structural models based on polygonal random fields.

### **Significant CENS-developed Databases and Repositories**

The previous studies would not have been possible without a stable base of well-documented data, from both short-term experiments and as well as long-term stable deployments. The development of effective fault detection methods, for example, requires detailed information from a variety of deployments. For each we ask: What equipment was involved? Were there known issues with certain observations or ranges of observations? Are other measurements available from nearby devices or from other instrumentation that could provide useful context? In this section we document three CENS-developed databases and repositories that provide uniform access to CENS data and research products. These projects have begun to converge, enabling CENS data, deployment records, and publications to be linked together.

#### *CENS Deployment Center (CENSDC)*

The CENS Deployment Center (CENSDC) is a database for researchers to document their deployment activities through the creation of pre-deployment plans and post-deployment feedback. CENSDC encourages researchers to

capture their experiences both in and out of the field to help inform future deployments. In so doing, we cull tacit information about equipment setups, deployment locations and field preparations, all of which play a critical role in data collection techniques. CENSDC designers have worked closely with CENS researchers as they plan for and ultimately execute detailed sensing experiments. Research is underway on a handheld mobile device for data collection in the field, from which data will flow into CENSDC and Sensorbase.

#### *Sensorbase*

For several years now, Sensorbase has provided a database backbone infrastructure for CENS. It provides a means to store all sensor data gathered in the field in a user-friendly environment (with a web interface for interactive use and a collection of APIs for programmatic access via SOAP and RESTful services). Sensorbase is built on a set of data permissions and privacy constraints that allow users to easily manage how their data are shared with others. This year, we completed a major code integration effort that ended with the creation of a central repository for distributing Sensorbase source code. In addition, we migrated our operation to a new, vastly more powerful database server and made public a host of new services for data upload and access.

#### **Studying the data lifecycle: Watching the watchers**

Finally, we present systematic studies of how data are collected, stored, shared and reused within CENS. This work takes place in a larger context of research into the data lifecycle and the design of effective cyberinfrastructure to support interdisciplinary collaborations.

#### *Towards a Virtual Organization for Data Cyberinfrastructure*

As part of our overarching research agenda to understand CENS as a collaboratory and the cyberinfrastructure (CI) necessary to support such a collaboratory, we established a CI Virtual Observatory for the study of data, data analysis, and visualization. In this one-year Small Grant for Exploratory Research, we developed a research instrument and sampling method to probe, in a consistent fashion, several CI projects, including CENS. The instrument is being used by collaborators at several institutions, each of which obtained IRB approval. Ultimately, we see this work providing guidance as to what constitutes a viable cyberinfrastructure.

#### *Object Reuse and Exchange; RDF data modeling*

This project concerns the design and development of tools that allow efficient reuse and exchange of information objects resulting from embedded sensing research applications. We are experimenting with the Open Archive Initiative's Object Reuse and Exchange protocol (OAI-ORE) to describe, publish and share aggregations of information objects produced at different stages of the scientific lifecycle in environmental sensing research. We are developing a sensor-specific vocabulary to describe the relationships among these information objects, using the Resource Description Framework data model to capture a) information about deployments (stored in the CENS Deployment Center), b) sensor data (stored in [Sensorbase.org](http://Sensorbase.org)) and c) scientific publications (stored in the California Digital Library's eScholarship repository). Although these data are organized and managed as separate entities in disjointed data archives, we view them as building blocks of the same scholarly production chain. Our present work is aimed at weaving together these resources using the OAI-ORE data model.

## SDP 01 Anomaly Detection

### SDP 01.1 People

- Principal Investigator: Leana Golubchik, Ramesh Govindan
- Faculty: Leana Golubchik, USC, Computer Science and jointly Electrical Engineering-Systems, Ramesh Govindan, USC, Computer Science, Professor
- Graduate Students: Abhishek Sharma, USC, Computer Science, Yuan Yao, USC, Electrical Engineering-Systems

### SDP 01.2 Overview

Anomaly detection is a difficult problem, even when the scope of the problem is reduced to a specific application. We believe that a robust (and more general) anomaly detection system will have to rely on multiple approaches simultaneously. And, that the appropriate interaction between the multiple approaches is an important ingredient in accurate anomaly detection. Thus, the broad goal of our project is to explore a variety of anomaly detection techniques as well as appropriate methods for combining them in a meaningful manner. To focus our project, we first began with a specific type of an anomaly, namely faults in sensor data, as described next. We are now expanding our efforts towards more general anomaly detection, and experimenting with these methods in the context of both, fault detection as well as anomaly detection.

Various sensor network measurement studies have reported instances of transient faults in sensor readings. In this work, we seek to answer a simple question: How often are such faults observed in real deployments? To do this, previously we reported on three qualitatively different classes of fault detection methods: rule-based methods, which leverage domain knowledge to develop heuristic rules for detecting and identifying faults, estimation methods, which predict “normal” sensor behavior by leveraging sensor correlation, flagging anomalous sensor readings as faults, and learning-based methods, which are trained to statistically identify classes of faults. We continued our efforts this year in exploring (a) time series based methods and (b) surprise-based method, which are described in detail below.

We apply our techniques to real-world sensor data sets and find that the prevalence of faults as well as their type varies with data sets. Our work is a step towards automated on-line fault detection and classification as well as more general anomaly detection.

### SDP 01.3 Approach

In our work thus far, we have focused on a small set of sensor faults that have been observed in real deployments: single-sample spikes in sensor readings (we call these **SHORT** faults), longer duration noisy readings (**NOISE** faults), and anomalous constant offset readings (**CONSTANT** faults). Figure 1 displays these faults in sensor measurements collected during different sensor network deployments.

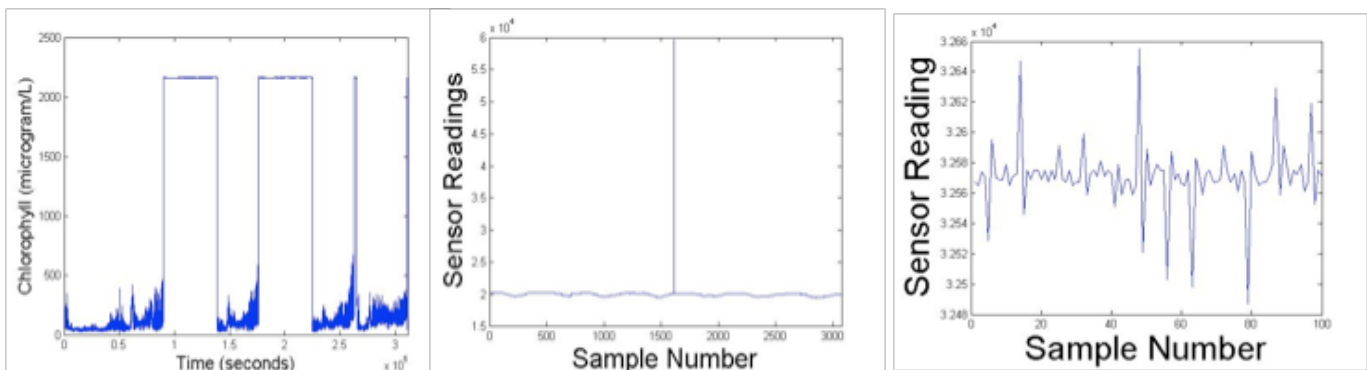


Figure 1: CONSTANT fault (left), SHORT fault (Middle), NOISE fault (right)

As already noted, we have explored a number of approaches to detecting faults in sensor networks, some of which have been reported in the previous report. In this report we focus on our recent experience with time-series and surprise-based methods.

We now give a brief description of a time series analysis based method. Time series analysis is a popular technique for analyzing periodically collected data. Specifically, the sensor reading can be viewed as a time series. These time series exhibit diurnal patterns as well as other (shorter) time scale temporal correlations. These temporal correlations can be exploited to construct a model of sensor measurements using time series analysis. To detect faults in a sensor measurements time series, we first forecast the sensor measurement at time based on our time series model. We then compute the difference between actual sensor measurement at time and its predicted value. We flag the measurement as faulty if this difference is above a certain threshold.

In addition to the time series method, we have been exploring a “surprise-base” method, which is briefly described as follows. Assume that we have a prior probabilistic estimation of the phenomenon being measured, say  $\mu$ . As the sensor measurements are collected, we construct a probabilistic model based on the obtained data, say  $\hat{\mu}$ . We refer to  $\hat{\mu}$  as the posterior model. This posterior model can be calculated using Bayesian inference. Intuitively, something “surprising” (or anomalous) is likely occurring if the difference between the prior and posterior model is sufficiently large. Thus we recursively compare and update these models to detect abnormal occurrences in sensor readings. We believe that this approach can be applied to fault detection as well as to more general anomaly detection.

#### SDP 01.4 System(s) Description and/or Experiments

Here we give a brief overview of our current experimental results. We study the performance of our methods by experimenting with data traces containing injected anomalies. The anomalies injected are of different intensities; this allows us to study the robustness of the proposed methods.

For time series analysis based methods, we use a multiplicative seasonal ARIMA time series model to model the

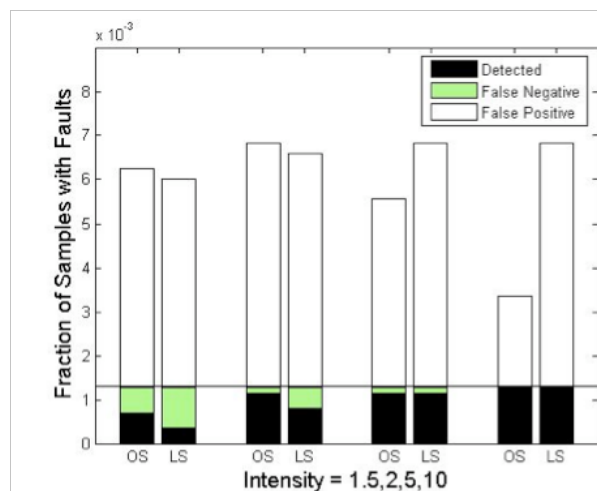


Figure 2: time-series based results

sensor measurements data sets. The parameters of the ARIMA model are obtained through a maximum likelihood estimation method. The accuracy of one step and  $L$  step ahead forecasting based detection methods are tested. We find that the performance of one step ahead forecasting based method is better in most cases in our experiments. The results are illustrated in Figure 2.

For the surprise-based method, we use normal distributions for both prior and posterior models. The parameters of the models can be calculated analytically from the collected measurements. The performance of the method is displayed in Figure 3. From this figure we see that the surprise-based method is able to detect most of the faults even when the fault intensity is relatively low. However, the false positive rate is relatively high. Exploration of this issue is part of our on-going efforts.

#### SDP 01.5 Accomplishments

In summary, during this reporting period, we:

- Developed and studied time-series and surprise-based methods to fault detection, and more generally to anomaly detection in sensor readings.

- Examined their efficacy using injected faults, partly to understand robustness characteristics of these methods..
- Examined with real-world data sets to understand behavior and accuracy characteristics of these methods.

### SDP 01.6 Future Directions

We believe that our work opens up new research directions in automated high-confidence fault detection, classification, data rectification, and so on. We plan to develop an online, automated sensor fault detection framework and integrate it with the existing sensor network architecture, such as TENET.

Moreover, we plan to continue expanding the project in the direction of more general anomaly detection, in the context of sensor systems and applications. Our goal is to explore a number of techniques and evaluate their effectiveness, while also focusing on the robustness of hybrid approaches. We plan to evaluate our approaches using the NAMOS data sets as well as other publicly available data sets.

One specific approach we plan to pursue is as follows. We note that another property of the real data sets is that the variance in the measurements at different times indicates different patterns as well as temporal correlations. For example, readings from light sensors exhibit larger variances during the day than at night. These behaviors is not captured by the ARIMA model. Thus our goal is to explore other time-series based models. For instance, we plan to pursue a GARCH model. The advantage of GARCH models is that they express the heterogeneity and variance correlations in time series. Thus, we expect this to be a promising future direction.

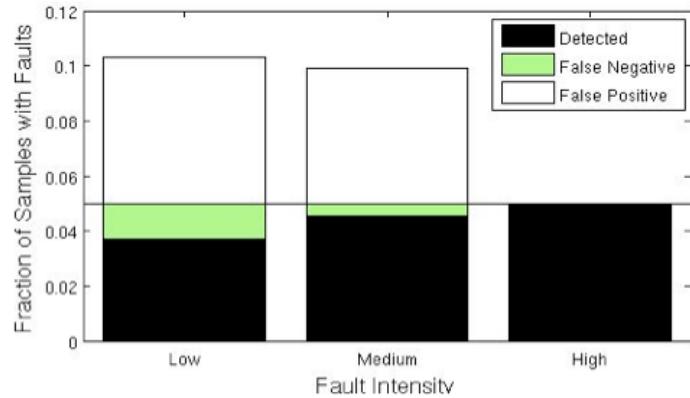


Figure 3: surprise-based results

## **SDP 02 Towards a Virtual Organization for Data Cyberinfrastructure**

### **SDP 02.1 People**

- Principal Investigator: Christine Borgman
- Faculty: Christine Borgman (UCLA, PI), Geoffrey Bowker (Santa Clara University, Co-PI), Thomas Finholt (University of Michigan, Co-PI)
- Researchers: Katie Vann (SCU)
- Graduate Students: Jillian Wallis (UCLA), David Fearon (UCLA) Mouly Kumaraswamy (UMich)
- Undergraduates: Victor Quintanar-Zilinskas (SCU)

### **SDP 02.2 Overview**

Our research work is primarily directed towards the design and development of tools and to allow efficient and durable management of data coming from embedded sensor networks. CENS, being an inter-disciplinary and multi-institutional center of innovation, is a convenient testbed to develop an integrated data management model as it fully encompasses the richness, heterogeneity and quantitative extent of current sensor network observations and their conceptual fragmentation across different academic groups and departments. Our research group is especially concerned with scientists' and engineers' data practices and their implications for science and education. Most of the group's work is devoted to research on data sharing, usability, preservation, accessibility and interoperability.

As part of our on-going research agenda to understand CENS as a collaboratory and the cyberinfrastructure necessary to support such a collaboratory we have joined with colleagues from University of Michigan, Santa Clara University, and Georgetown University to develop a comparative study of CENS and other collaboratories. Our long-term goal is to establish a CI Virtual Observatory for the study of data, data analysis, and visualization. Our short-term goal (in the time period of a one-year Small Grant for Exploratory Research (SGER)) is to conduct the background research to develop such a virtual organization, to identify the most effective models for a virtual organization, and to identify the research questions about data, data analysis, and visualization that must be addressed to construct a viable cyberinfrastructure. The promised outcomes of our work are:

- To produce fundamental social science about infrastructure development for virtual organizations and for data;
- To create a functioning virtual organization with a specific set of goals to study data practice and data policy;
- To identify methods for formative evaluation of CI efforts on data, data analysis, and visualization that will be enriched by continual comparisons between projects.

### **SDP 02.3 Approach**

Our virtual organization began with the writing of the SGER proposal, and has been extended in two convergent directions. First, we have conducted the initial stages of the project (literature review and development of methods and instruments) via distributed collaboration technologies that link the three university sites (UCLA, University of Michigan, Santa Clara University). Our use of technologies such as CTOOLS (a University of Michigan implementation of SAKAI), DimDim (web conferencing tools), audio conference services, videoconferencing (iChat and Skype), Google Docs, and Basecamp (a low-cost commercial site for sharing files, sending messages, whiteboard space) have enabled us to be participant-observers in our own collaboratory. The content of these distributed meetings is divided between administrative tasks and shared reading and discussion of relevant research. While each of these tools has strengths and weaknesses, Basecamp and audioconferencing are the most robust. We will continue to work with these tools as they mature. Recently, Thomson-Reuters made us a gift of Endnote licenses for use in testing its robustness for distributed collaboration. We have now settled into our chosen products and use them regularly for communications and to add items to the communal store.

The second direction is to extend our virtual organization through larger and longer grants to carry this research beyond the SGER award. We used the same distributed technologies noted above to develop these collaborations, adding more investigators from the same universities. Two of the proposals we submitted in early 2008 received funding. These are (1) HSD AOC – *Monitoring, Modeling & Memory: Dynamics of Data and Knowledge in Scientific Cyberinfrastructures* (Paul N. Edwards, UM, PI; Co-PIs Borgman, UCLA; Bowker, SCU; Thomas Finholt, UM; Steven Jackson, UM; David Ribes, Georgetown; Susan Leigh Star, SCU) and (2) *VOSS Mapping Uptake of VO Technologies* (Finholt, UM, PI; Co-PIs Jackson, UM; Ribes, Georgetown). These grant writing activities, and subsequently the awards, lead to the establishment of regular meetings between collaborating sites, both in-person and virtual. We held a 2-day in-person meeting of the investigators and graduate students of these three projects in December at Santa Clara University (with 2 participants joining by videoconference).

#### **SDP 02.4 System(s) Description and/or Experiments**

Part of the work stipulated in this research project was to administer a small program of research on data practices across multiple cyberinfrastructure projects. Part of our collaborative effort has been the development of a shared research instrument, with questions of interest to all of the investigators, and which capture a baseline of data to motivate future rounds of interviews. This instrument also can serve as exploratory research for the longer-term grants to follow.

Once our instrument was agreed upon, all institutions applied for and received full IRB approval. IRB approvals were received in November and December across the three participating universities. While awaiting approval of our instruments, we developed a shared sampling method based on centrality measures of participants in our research sites. Since we received our administrative approval we have contacted our sampled participants and begun interviews. Of the 40 total interviews to be conducted across the 3 sites, the majority of interviews have been conducted. We are now in the midst of data transcription and analysis.

#### **SDP 02.5 Accomplishments**

- Developed a cross-institutional research instrument and method
- Developed a sampling method based on author centrality measures and applied it to our research populations
- Obtained IRB approval of our study at all sites
- Invited participants and collected interview data
- Developed a shared method for data analysis
- Initiated data analysis

#### **SDP 02.6 Future Directions**

Over the next year we need to complete data collection, transcription, and analysis. Additionally we need to demonstrate the viability of this method for use within a cascading research organization that would allow future research to scale to meet the needs of both our growing virtual organization and the growing cyberinfrastructure projects we are studying.

#### **SDP 02.7 External Research Partnerships**

Anita Borg Institute (current)

This project is funded by NSF award #OCI-0750529, "Towards a Virtual Organization for Data Cyberinfrastructure". Christine L. Borgman UCLA, PI; Geoffrey Bowker, Santa Clara University, Co-PI; Thomas Finholt, University of Michigan, Co-PI. Start date: 2/1/08; End date: 1/31/09.

## **SDP 03 CENS Deployment Center (CENSDC)**

### **SDP 03.1 People**

- Principle Investigator: Christine Borgman
- Faculty: Christine Borgman
- Staff: Mike Taggart
- Graduate Students: Matthew Mayernik
- Undergraduates: Erick Romero

### **SDP 03.1 Overview**

CENS researchers are developing flexible wireless sensing technologies that can be used in a variety of scientific applications. These technologies are used to produce valuable scientific data. CENS data are largely collected on real-world deployments where sensing systems are deployed in particular locations where phenomena of scientific interest exist. CENS deployments are highly variable. Researchers cannot fully predict what they will encounter when they reach a field location. There may be unexpected temperature fluctuations, excess moisture or dust, or unpredictable flora and fauna, all of which affect the functionality of both equipment and people. Researchers have deployment goals and objectives prior to going out in the field, but decisions have to be made on site about where, when, and how to deploy equipment and sensors based on local conditions and the state of equipment. As a deployment proceeds, researchers adjust their activities in response to the situation, to time constraints, and to the available sensors and power. Trade-offs between these factors are made in real time, affecting the data that are captured during a deployment and the ways that they are interpreted post-deployment.

As CENS researchers participate in deployments, they build up knowledge about potential problems they may encounter and how to solve them. Community knowledge of deployment best practices is a valuable asset for CENS. Our central research questions are how to facilitate deployment knowledge transfer in the collaborative CENS research setting, and how information about deployment activities can be leveraged to describe CENS data.

We have expanded CENSDC development efforts in two main ways, first through the development of a multi-site deployment module, and second to enable researchers to collect information about deployment activities while deployments are taking place. Early user feedback suggested that the inability access the internet while in field locations created a significant gap in our system. In our work this year, we have approached this gap in two ways: 1. developing handheld methods for collecting deployment information as deployments are taking place, and 2. interfacing with streaming data. This led to a successful proposal to Microsoft Research to develop an application for a handheld device for in-field use, which is discussed more in later sections of this report, and in greater detail in a separate report.

### **SDP 03.2 Approach**

The CENS Deployment Center (CENSDC) was designed to leverage CENS deployment knowledge by providing a central location for researchers to document deployment activities through the creation of pre-deployment plans and post-deployment feedback/notes. By allowing users to describe their deployment experiences, including lessons learned, troubleshooting techniques, and provide guidance for future deployments, we are attempting to capture the tacit knowledge about equipment setups, deployment locations, and field preparations that play a critical role in data collection techniques. As CENS technologies mature and current researchers gain deployment experience, new students face a steeper learning curve when joining a project.

A parallel goal of the CENSDC is to add value to CENS data by providing a source of descriptive and contextual information surrounding the data collection. CENS provides a great cyberinfrastructure test case on data reuse. The data from CENS deployments are spatially and temporally located, having scientific value both to immediate research questions and long-term longitudinal studies, and are therefore irreproducible. With any research

endeavor, understanding the context of data collection is critical to the ultimate evaluation and interpretation of results. The challenge of capturing data collection context is particularly difficult when data are collected on highly variable and unpredictable real-world deployments. The process of collecting and articulating information about deployments in the CENSDC involves valuable reflection on research methods, sample selection, and troubleshooting techniques. This information can assist researchers in writing papers, proposals, and reviews, as well as in maintaining their data and leveraging them for reuse by others.

### SDP 03.3 System(s) Description and/or Experiments

System development in the past year has primarily focused on implementing a module that records information about multi-site deployments. The CENSDC system was designed to address the needs of singular campaign-style deployments. We have expanded the system to incorporate a multi-site deployment module uses a parent-child model where a number of individual deployment sites are constituent parts of a single deployment. The development of the multi-site module was undertaken to specifically address the complex deployments performed by the CENS seismic group. A multi-site module has been implemented and is being used to collect regular updates from the seismic deployment.

The second main development effort has focused on interfacing with regular status updates from ongoing deployments. As the Peru seismic deployment is progressing, more and more data are being streamed to UCLA, including various kinds of technical data, along with the seismic data itself. The technical data provide various measures of the health of the seismic stations, as well as of the wireless networks that connect them to each other and to the internet. These measures are meant to enable members of the seismic team to characterize the status of each station, and to identify problems as they arise. We have created a Google Maps interface that pulls data from the seismic team database to map the quality of the wireless links between stations along the deployment transect. The map interface is shown in Figure 1. The link quality data on which the map is based are updated every day, enabling members of the seismic team to view the current status of the station-to-station wireless links. The mapping feature is being evaluated with members of the seismic team, and our intension is to expand the feature to show other kinds of data. Next steps will include mapping data flows between stations.

### SDP 03.4 Accomplishments

Members of the CENSDC team have taken part in five deployments or research trips in the past year. CENS research takes on many forms. Each research team has their own ways of performing field research, and collaborations between research teams add to the complexity. Sensor network deployments occur in different types, reflecting this diversity of research methods, and fit within the larger scope of research projects. Deployment planning and organization also reflects the research diversity within CENS, as each team coordinates and organizes within the dynamic of their own group.

In expanding use of the CENSDC, we



Figure 1: Map of Wireless Link Success in a subsection of the Peru seismic

have found that the system is most suited for projects that conduct repeated or long-term data collection deployments. For projects with repeated deployments, equipment and data collection processes are often similar from deployment to deployment, but the specific differences between deployments are important. The CENSDC “make like” function enables researchers to create new deployment records from past deployment information, illustrating the continuity between deployments. For long-term sensor deployments, monitoring and maintenance of deployment equipment and sites are critical throughout the lifetime of the project. Logs of these activities in the CENSDC serve as data annotation, giving an overall picture of the problems encountered and solutions during the deployment. Additionally, the most active users of the CENSDC have been involved with deployments that take place in remote locations, such as in Peru and Bangladesh, as the organizational requirements for performing remote deployments are much greater than for local deployments. Members of local deployments can adapt to changing plans or deployment needs with more flexibility.

To connect deployment information in the CENSDC with the data that results from CENS deployments, we are linking from deployment records to online data sources. Most deployments in the CENSDC do not have associated projects in Sensorbase. None of the recent NAMOS, Bangladesh, Merced River, or Seismic data are held in Sensorbase. However, some of these projects have data available online separate from Sensorbase, such as the Merced River and NAMOS projects. We are linking to these additional CENS data sources where possible, and have plans to create more direct Sensorbase-CENSDC crossovers, as detailed in the next section.

### **SDP 03.5 Future directions**

Future work on the CENSDC will focus on increasing the visibility and use of the system by creating Sensorbase-CENSDC crossover functionality, performing user studies and evaluations of the CENSDC website and functionalities, and continuing to work on collecting and displaying information about deployments as they are taking place.

One of the goals of CENS is to make high quality, well documented data sets from CENS deployments available to other researchers. Sensorbase provides researchers with a platform for data sharing, but as mentioned above, Sensorbase projects typically do not have much description or documentation associated with them. Detailed documentation is not necessary for all Sensorbase projects, as for various reasons many projects are not intended to be used in contexts outside of their immediate use. But for projects where the data owners would like to share their data, or at least make it available to others, additional description may be beneficial. We have been in discussion with the Sensorbase team regarding ways that some CENSDC features might be integrated into Sensorbase, specifically to enable users to give more information about the dates and locations that data were collected, the people involved in the project, and equipment used to collect data. Additionally, we plan on enabling CENSDC users to create Sensorbase projects when creating new deployment plans, as a means to reduce redundancies across the two systems.

A further benefit to building bridges between the two systems is that Sensorbase projects are not indexed by Google or other search engines. CENSDC project pages that describe Sensorbase data would provide more visibility for CENS data, as the CENSDC pages are indexed by the major search engines. CENSDC descriptions of Sensorbase data will not replace the need for researchers to personally discuss the specifics of sharing data, but they can be a way for outside users to quickly assess the potential usefulness of CENS data, and find out who to contact for more information. In this sense the CENSDC pages would serve as the public face to privately held data.

In parallel with the integration efforts, we will conduct user studies of the existing CENSDC system. User studies will focus on two main aspects of the system. The first aspect is collaborative knowledge transfer. We will evaluate the utility of the CENSDC in helping new members of a research team to get up to speed on the kinds of deployment problems and issues to expect, as well as the utility of the CENSDC in keeping current team members updated on the status of a deployment. The second aspect we will evaluate is the utility of the information captured in the CENSDC for data description and discovery. Data is often not self-describing, in particular Sensorbase projects typically do not have much description associated with them. Many different kinds of metadata schemas are available to describe different kinds of data, but are often difficult to understand and use. Our goal is to enable

researchers to collect deployment information in the CENSDC that can be used to supplement existing metadata, or when metadata is absent to serve as a data description that can be used to share and discover data resources. We will evaluate the effectiveness of CENSDC deployment information in describing CENS data through surveys and informal interviews with CENS researchers.

Further deployment and use of CENSDC will focus on the types of deployments described that we have found to be most suited to the system architecture, those with similar and repeated deployments and those with long term sensor installations. We will also look at ways that we can expand the integration of real time deployment displays like we have created for the seismic deployment. The need to know the status of existing equipment installations is not unique to the seismic deployment, other deployments have similar needs. Developing similar deployment displays will be based on the needs of individual deployments, with the goal of producing generalizable best practices for displaying deployment status information.

Additional future work centers on the development and deployment of the handheld application for deployment data collection. This work, described in more detail in a separate annual report, feeds into the CENSDC development by providing a mechanism for collecting information about field activities as they are occurring. The lack of in-field access was identified as a significant gap in our existing system, and we hope to address it through a handheld-based data collection and note-taking tool.

## **SDP 04 Monitoring, Modeling, & Memory: Dynamics of Data and Knowledge in Scientific Cyberinfrastructure**

### **SDP 04.1 People**

- Principal Investigator: Paul Edwards (University of Michigan), Christine Borgman (UCLA)
- Faculty: Paul Edwards (UMich, PI), Christine Borgman (UCLA, Co-PI), Geof Bowker (Santa Clara, Co-PI), Steven Jackson (UMich, SP), David Ribes (Georgetown, SP)
- Graduate Students: Jillian Wallis (UCLA), David Fearon (UCLA)

### **SDP 04.2 Overview**

As framed in the NSF Cyberinfrastructure Vision report, scientific data can be key contributors to human progress, learning, and discovery. But present reality falls short of this ambition: despite large and growing investments, scientific data are not widely available for reuse; data sharing between researchers and disciplines is limited; and standardized practices for data access, curation, and provenance remain weak or ineffective. Too little is yet known about the dynamics of data and knowledge in transdisciplinary scientific cyberinfrastructures (CI). How are data generated, stored, and shared across teams, institutions, and disciplines? What factors make data robust and trustworthy in distributed transdisciplinary research environments? How do individual data points grow into stable, usable, and innovative knowledge? These are neither matters of faith nor simple technical fixes. This project begins to fill that gap via empirical research.

Advanced cyberinfrastructure challenges and extends scientific practice in three crucial ways. First, large numbers of automatic sensors monitor subjects of interest, producing massive volumes of digitized data. Second, computational models drive data collection, prediction, experimentation, and decision-making in a growing number of fields. Third, increasingly vast data resources (scientific memory) are collectively available, though often distributed across thousands of research sites, institutions, and communities. If CI-enabled science is to deliver on its transformative potential, the dynamics of data and knowledge production (old and new) must be understood, and criteria for success and best practices established.

This project investigates practices of monitoring, modeling, and memory across four leading CI projects targeting three critical domain areas: ecology and environment (LTER and CENS); hydrology and water management (the WATERS network); and earth systems science (ESMF), united through their relevance to climate change concerns. Our project sites: a) reflect the ‘state of the art’ in current CI investment; b) support comparative analysis through an appropriate mix of shared and divergent data challenges; c) represent critical domain areas in which project payoffs will have immediate and important consequences; and d) build on the research team’s own histories of collaboration and domain expertise.

Methodologically, the project develops an innovative combination of distributed ethnography, collaborative history, and multimodal network analysis in large-team settings – creating a model for future research of this sort.

### **SDP 04.3 Approach**

This project will expand understanding and improve performance of the already substantial investments in cyberinfrastructure made by NSF and other funders. To this end, along with original research findings (made available on open access terms through venues such as the UC’s eScholarship or Michigan’s DeepBlue repository), we will produce a handbook of CI Best Practices meant to guide data practices and collaborative coordination among existing and future CI projects. Working with our project and outreach partners, our research will lay groundwork for an inclusive, theoretically rich, and practically engaged social science of cyberinfrastructure.

Our project will make immediate contributions to data practice and collaborative dynamics within the four projects under study. More broadly, it will help shape and inform science, education, and policy-making within the critical domain areas of ecology, water, and climate science. It will enhance infrastructure for learning by making research data more widely available for instruction at the K-16 through graduate levels. Through our outreach partners, we

will explore modes and patterns of exclusion embedded in existing cyberinfrastructure dynamics, and develop more robust analytic capacities for mapping and remedying these patterns in future through the design and redesign of existing and emergent cyberinfrastructure. Beyond its theoretical contributions, our project will significantly improve both practical implementation and broad-based participation within emergent cyberinfrastructure. Key, unanswered research questions for the CI vision therefore include:

- How do participants from one disciplinary community make sense of data produced under the very different procedures and background assumptions of another?
- What kinds of knowledge do scientists require to make effective use of “foreign” data?
- What factors most influence scientists’ trust in data and data-sharing tools, as collaborative webs expand and their first-hand knowledge recedes?
- How, and how much, can designers, managers, scientific users, and social scientists work together to create the social, organizational, and institutional prerequisites for successful large-scale collaborative work?

#### **SDP 04.4 System(s) Description and/or Experiments**

To answer these questions, we have begun (October, 2008) a three-year comparative study of four major cyberinfrastructure projects. We chose these projects because each involves 10-100 participating institutions, seeks cross-disciplinary collaboration through cyberinfrastructure, spans multiple temporal and spatial scales, and engages central issues of monitoring, modeling, and scientific memory. Further, while the individual projects involve separate domain sciences, all relate centrally to environmental change. In the long run, they might potentially be linked in an even larger infrastructure. We will analyze each project using a range of methods from oral history to ethnography and relational-dynamics mapping. Simultaneously, our research team will compare the four projects in an iterative cycle, leading to outcomes such as a “CI Best Practices” manual of lessons learned for large-scale CI projects.

#### **SDP 04.5 Accomplishments**

- Participated in a kick-off meeting where we shared our goals and sites with one another
- Developed a shared research protocol, including instrument, tools for collaborative data collection and analysis, as well as the rules to guide consistent application and use of these tools.
- Currently collecting baseline data from participants at our research testbeds

#### **SDP 04.6 Future Directions**

During the next year we will begin the following research initiatives:

- Conduct semi-structured interviews
- Perform participant observations
- Collect and review documentation from the early and pre-histories of each project
- Map and compare biographical trajectories of key project personnel
- Conduct oral history interviews with key project personnel
- Compile, evaluate, and map available quantitative project data
- Identify boundary objects within and across the target projects
- Construct maps of relational dynamics and relational clusters

## **SDP 05 eScholarship Repository**

### **SDP 05.1 People**

- Principal Investigator: Christine Borgman
- Faculty: Christine Borgman (UCLA)
- Staff: Xiao-Mai Vo (UCLA), Jeff Goldman (UCLA), David Avery (UCLA)
- Graduate Students: Alberto Pepe (UCLA), Jillian Wallis (UCLA), Tommy Keswick (UCLA)

### **SDP 05.2 Overview**

Institutional repositories are often seen as the solution—or at least a step in the right direction—for a number of different problems facing the academic world. Problems such as the scholarly communication crisis that have resulted from rapidly increasing journal subscription prices to the ability of libraries to house and preserve copies of journals that have gone electronic can all be addressed by institutional repositories. Repositories, because of their web-based nature, are also claimed to bring additional benefits to those authors who deposit in them such as increased citation rates and new metrics for assessing use of materials (e.g., download statistics and page hits). Institutional repositories also fit with the open access agenda, specifically utilizing Open Archive Initiative standards to support dissemination of bibliographic data to web-harvesters. By the nature of being “institutional”, institutional repositories have behind them many resources that disciplinary repositories may not have. Name recognition, longevity, and funding sources are among the institutional advantages when compared to subject repositories that may be scattered across many different locations.

### **SDP 05.3 Approach**

We are building an architecture for data integrity and quality in wireless sensing systems. The eScholarship Repository is part of a larger data ecology along with Sensorbase.org, CENS Deployment Center, and other realtime data integrity initiatives such as Confidence. Each of these systems captures part of the data context, and linked together overcome the limitations of isolated systems, creating a robust description for each dataset thereby supporting reuse.

### **SDP 05.4 System Description**

CENS has maintained a web-accessible bibliographic database of publications since its inception, but this system has not scaled well to meet the needs of researchers or aged well in light of web 2.0 functionalities. The eScholarship Repository is an institutional repository maintained by the UC System, which allows schools, departments, and research centers to deposit their documents. The repository provides an array of access, distribution, maintenance, and curation services .

The metadata in the repository is more bibliographic in nature, and more expressive than our existing bibliographic database, which allow for more sophisticated discovery tools, such as filtering by author and subject. When we made the move to eScholarship we made the commitment to clean each record in order to restore lost functionality. Cleaning each record includes but is not limited to: re-entering author information; entering a richer description of the resource including subject, abstract, journal or conference information; researching publisher copyright information and tracking down the appropriate version of the item to be posted in accordance with the copyright statement of each publisher. Once clean, the CENS eScholarship Repository will allow us to perform social network analysis based on collaboration and co-authorship to augment our research findings.

### **SDP 05.5 Accomplishments**

- New fields for bibliographic description were identified and requested from bePress (the service provider)
- With the help of two CENS summer interns, materials from the '07-'08 year were cleaned and uploaded to the eScholarship Repository

- We have developed a plan for removing, cleaning, and re-uploading the existing data to take advantage of fields that were added at our request, and to clean up the records that have not yet had authors split out
- We have liaised with the California Digital Library, who are responsible for the eScholarship Repository, with regards to features we would like to have added to their front-end redesign

#### **SDP 05.6 Future Directions**

- We will continue to add items to the repository
- We will continue to assist authors in the repositing process
- We will follow the plan for cleaning the rest of the records

## SDP 06 Handheld application for mobile data collection in the field

### SDP 06.1 People

- Principal Investigator: Christine Borgman
- Faculty: Christine Borgman – UCLA, Information Studies
- Graduate Students: Matthew Mayernik, Alberto Pepe - UCLA, Information Studies
- Undergraduates: Erick Romero – UCLA, Computer Science

### SDP 06.2 Overview

Environmental research data is primarily collected outside of lab settings in real world field locations. Field locations are highly variable; researchers adjust their in-field activities in response to the immediate situations, such as inclement weather, time constraints, and equipment breakdowns. Field deployment activities affect the data that are captured and the ways that they are interpreted post-deployment. We are developing the CENS Deployment Center (CENSDC) as one step to address the flexible and unpredictable nature of field research. The CENSDC is a web-based system for the collection of deployment information, such as equipment lists, field notes, and suggestions for future deployments. However, the web-based approach is limited by the lack of internet conductivity in most field settings. We believe that this problem could be addressed through the development of a handheld data collection, note-taking and microblogging annotation application for use during field-based environmental research. This report outlines our work in developing a handheld application for simple data collection and note-taking in the field.

### SDP 06.3 Approach

This application builds on CENS expertise in development for handheld devices, particularly the EcoPDA and Campaignr projects. We plan to extend their current work to develop a generic module for deployment use that will supplement and interoperate with our CENSDC system, thus testing both interoperability and mobility issues in ecological field research. CENS researchers engaged in field research normally work in small teams in specific locales using ad-hoc methods and heterogeneous data sources, and adjust research activities on the fly. The goal of our research is to develop tools and applications that facilitate the collection of contextual data about CENS research activities and similar field-based sensor network research. Our development principles for the mobile application are the following:

- *Facilitate collaboration and sharing:* cyberinfrastructure initiatives promise increased collaboration across distances and increased sharing of informational resources. Neither of these can happen without standardized ways of describing research methods and products. Our application will utilize both existing standards, such as the Environmental Metadata Language and new methods, such as RDF vocabularies, to facilitate collaborations and sharing both within and across projects.
- *Design for activity:* Ecological researchers engage in various kinds of activities while in the field. Our application will be designed with an understanding of these activities, and the contexts in which they are to be performed.
- *Facilitate emergent research dynamics:* Field ecology is a highly emergent science. As researchers spend more time in a particular field location, they become more aware of the variables that are relevant to their study and adjust their research activities accordingly. This emergent organization of activity leads to particular research dynamics and patterns of activity. Our application must facilitate the creation and communication of emergent field dynamics within collaborative research.

These principles lead to our development goals. We want the application to enable researchers to:

- *Specify data collection protocols:* Researchers collect highly varied kinds of data. The application must allow researchers to specify their own collection procedures, and customize the data collection interface to those procedures.

- *Collect repeatable data*: Researchers often repeat data collection procedures, both to repeat and augment prior experiments. Our system must allow them to re-use data collection protocols that they have already created and used.
- *Perform their field activity as fast with the tool as without the tool*: Time spent in the field is precious. If our system noticeably slows researchers down, its usefulness is greatly diminished.
- *Easily integrate the field data collected with the tool into their existing data collections*: Researchers in most sciences are already facing challenges in organizing and using digital data resources. Our application is intended to be one step in mitigating this problem, thus our system must integrate with existing data and data tools.
- *Produce data process descriptions (collection procedures, annotations) that “live with” the data*: Data are collected in various ways. Understanding and using data requires understanding of the way that data were collected. Providing additional information about the data collection processes that “lives with” data increases the usefulness, and consequently the value, of the data itself.

### SDP 06.4 System(s) Description and/or Experiments

The first main functionality of the application is data collection. The application will allow researchers to collect data in tabular form. More specifically, the data collection functionality will contain the following features:

- “Authoring” – the web interface will allow the user to create spreadsheets and data collection protocols before going out in the field on their personal computer. They can then upload the spreadsheets as XML files to the handheld device for use in the field. These data collection forms will be customizable, so that they can be changed, and re-usable, to facilitate repeatability.
- Import/export features to Excel, csv, Sensorbase

The second main feature of our application is the note-taking/microblogging functionality. “Microblogging” refers to the use of web services such as Flickr and Twitter to enable researchers to post pictures and short notes to the web from field locations. Both notes and photos can be posted as microblogs. Notes or photos can be associated with particular items (locations, equipment, serial numbers, tasks, people, pictures, etc.) as metadata. For example, when making a note, it will be associated with one or more aspects of the current data collection activity, including the current transect, current location, current person, current data point, etc.

### SDP 06.5 Accomplishments

Work on this project began in June of 2008, and to date has focused on doing background research, gathering user requirements, building models and specifications, and beginning prototype development. Figure 1 shows the concept map that was created as the requirement building and specification process. Another part of our background work has included identifying existing CENS code that we may reuse/customize to our project. We are using Windows Mobile 6 as the basis for our application, which allows us to build off the Windows Mobile-based EcoPDA project. For certain functionalities, we can also be able to build off of the code base of the CENS Campaignr project, which has recently been migrated to the Windows Mobile platform as well.

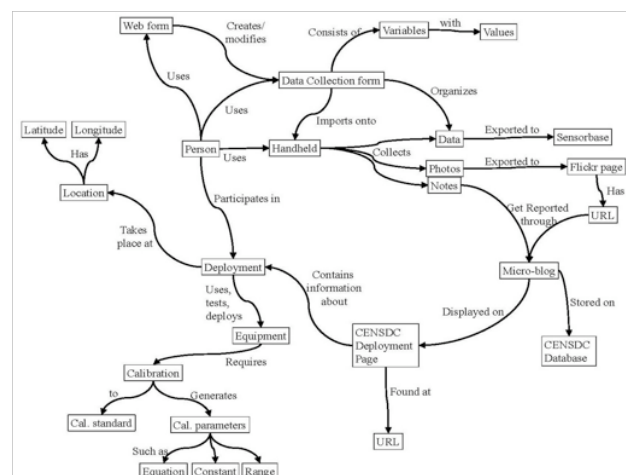


Figure 1 Concept Map

### **SDP 06.6 Future Directions**

Future work will focus on iterative development and testing of the system, first of prototypes and subsequently of fuller systems. Initial tests will be with the CENS beach monitoring group, with further tests to follow.

In the longer term, we would also like to build in additional functionalities that allow for more flexible and nuanced data collection than is possible with simple spreadsheet tables. These additional functionalities could include:

- Enable researchers to run simple statistical checks or models while in the field as a means of determining sample sizes, densities, etc.
- In-field data graphing or visualization
- Support multiple kinds of data collection procedures: transects, quadrats, time-series.

### **SDP 06.7 External Research Partnerships**

Microsoft Research (current)

## SDP 07 Object Reuse and Exchange; RDF data modeling

### SDP 07.1 People

- Principal Investigator: Alberto Pepe, Christine L. Borgman (Information Studies, UCLA)
- Faculty: Christine L. Borgman (Professor and Presidential Chair, Information Studies, UCLA)
- Graduate Students: Alberto Pepe, Matthew Mayernik, Jillian Wallis (Ph.D. students, Information Studies, UCLA)

### SDP 07.2 Overview

The research work presented here is primarily directed towards the design and development of tools to allow efficient reuse and exchange of information objects resulting from embedded sensor network research applications. We describe the utilization of the Open Archive Initiative's Object Reuse and Exchange protocol (OAI-ORE) to describe, publish and share aggregations of information objects produced at different stages of the scientific lifecycle in environmental sensing research. Moreover, we propose to develop a sensor-specific vocabulary to describe the relationships among these information objects, using the Resource Description Framework data model.

### SDP 07.3 Approach

The work presented here builds on previous research in which we developed a conceptual model of the CENS scientific lifecycle (discussed in [4]). This research has revealed that production of environmental sensing data involves continuous handling of heterogeneous types of information at various stages of a data life cycle, from data collection to data curation. We identified three major digital resources across the CENS data life cycle: a) information about deployments (stored in the CENS Deployment Center), b) sensor data (stored in Sensorbase.org) and c) scientific publications (stored in the California Digital Library's eScholarship repository). Although these data are organized and managed as separate entities in disjointed data archives, we speculate that they are all building blocks of the same scholarly production chain. Our present work is aimed at weaving together these resources using the OAI-ORE data model.

### SDP 07.4 System Description

In ongoing work, we are extending the notion of scientific lifecycle for ecological sensing data. In particular, we are proposing a technical system implementation of the OAI-ORE protocol to enable description of the processes and practices which take place in a typical CENS scientific lifecycle. Via this implementation, we plan to aggregate all information objects produced in the scientific lifecycle: scholarly publications, stored in the eScholarship repository, as well as information on which those publications are based, such as interim drafts, and contextual information, stored at the CENS Deployment Center and raw sensor datasets, stored at Sensorbase.org. Our speculation is that by linking these related information objects in meaningful aggregations can yield advantages to the scholarly communication process as a whole, as the relationships between these products of the scientific

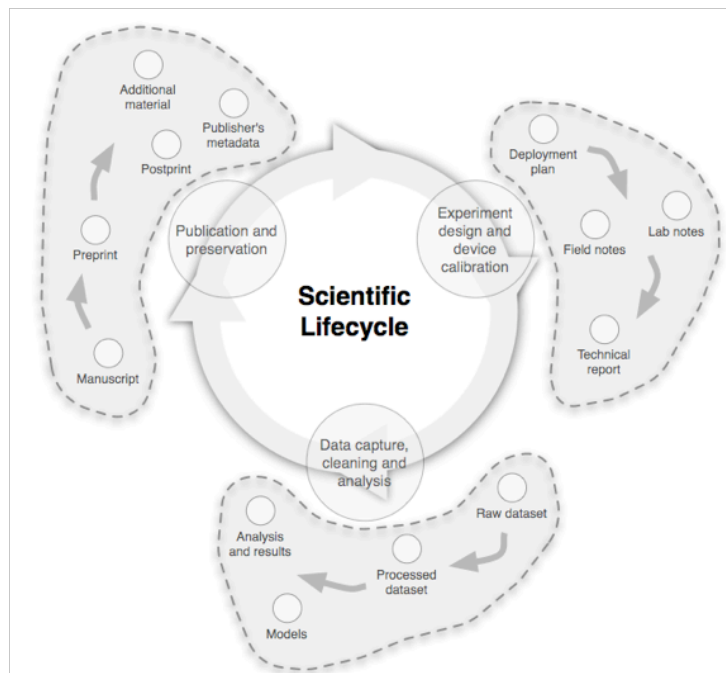


Figure 1. The data lifecycle of scientific research in environmental sensing [1]

lifecycle are preserved. Figure 1 depicts a conceptual enhanced version of the scientific lifecycle for a typical environmental sensing application.

### **SDP 07.5 Accomplishments**

Specific accomplishments during the reporting period:

- Developed and disseminated model of the CENS data lifecycle [4].
- Demonstrated the potential use of RDF to model information object relationships for network analytic applications [2,3,5,6]
- Thus far, awarded a Microsoft Research Gift to sponsor until the end of the 2009-2010 Academic Year both this project and the “Handheld data collection device” (led by Matthew Mayernik).

### **SDP 07.6 Future Directions**

In the coming year, we plan to produce and disseminate some OAI-ORE aggregations that represent concrete sensing applications at CENS, one in the field of seismology (the Southern Peru, formerly Middle American Subduction Experiment (MASE)) and one in the field of environmental science (the Networked Info-Mechanical Systems (NIMS) deployment in the San Joaquin River). These aggregations will be documented in a scholarly article, currently in preparation [1]. This will allow partner scholarly and scientific repositories to harvest aggregated information about CENS research in a structured format. In subsequent work, we plan to enhance these compound object descriptions by developing a sensor-specific vocabulary, using the Resource Description Framework data model. This will allow us to describe in detail both the relationships among aggregations and those between single data components within these aggregations.

### **SDP 07.7 External Research Partnerships**

- Microsoft Research. (Current) The Technical Computing Group of Microsoft Research supports Alberto Pepe’s participation in this project until the end of the 2009-2010 Academic Year. The funding is provided in the form of a gift.
- Los Alamos National Laboratory. (Current) The research involved in this project is often performed in collaboration with the Digital Library Research and Prototyping Team at the Los Alamos National Laboratory, New Mexico.

## SDP 08 Data integrity for TEOS

### SDP 08.1 People

- Principal Investigator: Prof. Mark Hansen
- Faculty: Mark Hansen, Statistics, UCLA
- Graduate Students: Sheela Nair, Statistics, UCLA

### SDP 08.2 Overview

A major issue that limits the widespread use of sensor networks is the quality of sensor data, which are compromised by various faults and anomalies. Over the past few years, the Data Integrity group discussed challenges and problems in ensuring that data collected from sensors are not corrupted by sensor faults. An ongoing goal has been to develop well-characterized fault models so that a system can determine whether anomalous sensor behavior is truly a fault. The paper, "Sensor Network Data Fault Types" (to appear in the *2009 Proceedings of the ACM Transactions on Sensor Networks*), identified a taxonomy of fault types as well as provided some ideas of features that can be used to model these classes of faults. This past year, work on fault detection has focused on two main areas. The first was identifying data features (e.g., transformations of the received data) that are useful for detecting each of the fault types previously identified. The second was studying the tradeoffs of having a fault detection system that adapts to changing sensor behavior in the presence of possible faults.

### SDP 08.3 Approach

We propose a signature-based fault detection system for identifying both intermittent faults as well as sensors which are experiencing persistent faults. This approach involves identifying a set of features that can be used to model normal behavior and faulty sensor behavior. Each sensor has its own signature of normal behavior, which allows for sensor-specific fault detection. Sensor signatures are periodically updated as new data are received which allows the signatures to adapt to changing sensor behavior over time and ensures the signatures are current. When a new observation is received, it is compared to the sensor signature and a fault signature. If it "looks" too much like a fault, the observation is flagged. We develop a general signature-based algorithm for use with any sensor types and deployment set-up, and the formulation allows a practitioner to incorporate as much deployment-specific information as is desired.

Our contributions have two main components. First, we identify data-dependent features that are useful for identifying the occurrence of faults and which combinations of features can characterize each of the different fault types. Different fault types identified in the taxonomy often have specific departures from the normal sensor behavior, and current research involves identifying features in which the signature or signal of a fault is very different from normal behavior. Features where this departure is most different from normal behavior will best discriminate between normal and faulty behavior. We explore the use of both model-based features and data-based features that make use of either spatial, temporal, or spatio-temporal information. Figure 1 below shows a stuck-at fault (when the sensor gets "stuck" at an incorrect value and reports that value for a number of successive observations. We look at temporal difference feature,  $Y(t) - Y(t-1)$ . Using this feature, the stuck-at fault is quite easy to characterize, as the signal is exactly zero for the duration of the fault.

As sensor behavior and the processes being studied can change over time, it is desirable for the model of normal sensor behavior to adapt to changes over time. A second contribution of our work is to provide methods for fault-tolerant parameter estimation. Our signature-based algorithm allows for signatures to "learn" changing sensor behavior over time. This means that parameter estimates or distributions of features must be updated online. When a non-parametric signature is used, the system must maintain (possibly) multivariate distributions of features must over time. The fact that this update is done in the presence of faults makes this a more difficult problem that has not received much attention in the fault detection or process monitoring literature.

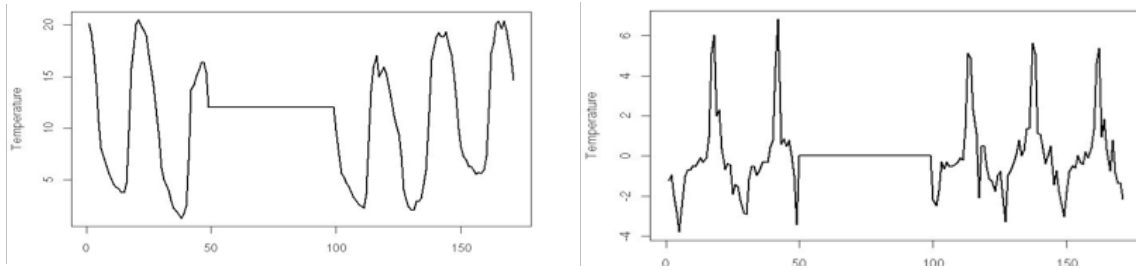


Figure 1: Left Panel: Stuck-at fault in AMARSS transect temperature sensor; Right Panel: fault signal in the simple temporal feature:  $Y(s,t) - Y(s,t-1)$

### SDP 08.4 System(s) Description and/or Experiments

Data collected by sensor networks often exhibits complex dependencies, over both time and space. We set up and conducted a simulation study to examine the effect of the space-time dependency structure in data on the detection performance of various features. To model data from sensor  $s$  at time  $t$ , we consider a general class of space-time models of the form:

$$Y(s,t) = m(s,t) + f[Y(s,t-1) - m(s,t-1)] + e(s,t),$$

where the errors  $e(s,t)$  are correlated across space.

Several temporal features that are frequently used in the fault detection literature for detecting deviations from expected behavior are investigated for their ability to detect sensor faults in our taxonomy. We also develop and study corresponding spatial and spatial-temporal analogs and evaluate the gain in detection power when spatial information is used. Using simulated data, we look at how the features are affected by various degrees of both temporal and spatial correlation.

Next we use these features to detect faults in data from an example deployment. We use the AMARSS transect at James Reserve and focus on the temperature sensors. There are 10 stations (nodes), each with four temperature sensors at different depths: one at the surface and three belowground. We evaluate the performance of non-parametric univariate signatures. We inject faults of different types in the real data to see how the algorithm performs at detecting known faults.

### SDP 08.5 Accomplishments

We have studied the detection capabilities of various temporal, spatial, and spatial-temporal statistics on various fault types. Simulations show that, when parameters are known (an assumption which is not practical in real life), there is a lot of gain in using spatial-based features in addition to purely temporal features (as seen in Figure 2).

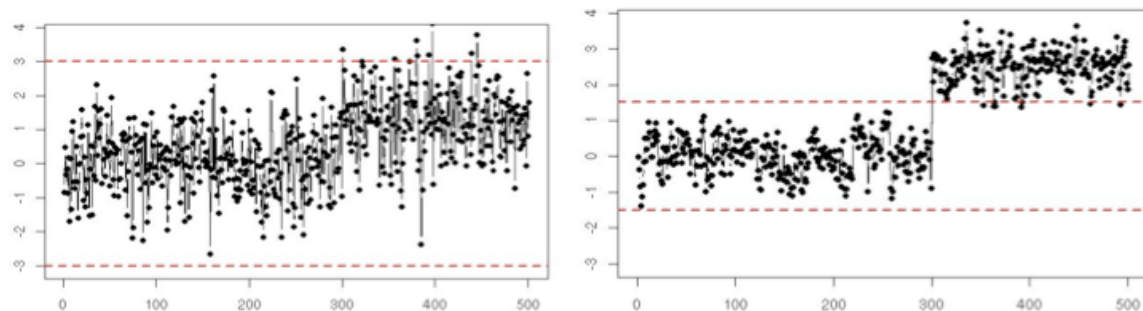


Figure 2: These plots show the signal of an abrupt jump in mean in two model-based univariate signatures. (Data are simulated.) The mean change occurs at time 800, and the red lines represent the thresholds for flagging a fault. Left Panel: Feature is the residual from temporal prediction; Right Panel: Feature is the residual from spatial prediction. The signal-to-noise ratio of the fault is much stronger in the spatial feature so signature using the spatial information will detect the fault much quicker.

Surprisingly, features which incorporate both spatial and temporal information are not as useful. However, when parameters are unknown, the spatial statistics often require estimation of more parameters than the temporal analogs because parameter estimates (e.g.,  $m(i,t)$  for many locations  $i$ ) at multiple locations are required. The variance of a feature's distribution increases with estimation of additional parameters. Therefore, the signal-to-noise ratio of the fault is reduced and the signatures using the features have decreased power to detect the fault.

### **SDP 08.6 Future Directions**

Many of the environmental processes studied in CENS deployments are subject to various fine-scale effects, such as patches of clouds or gusts of wind, that are difficult to predict or model. This is a main difficulty for many of the features we have studied thus far, as they are sensitive to any deviations from expected or predicted behavior. This results in a higher number of false positives from the fault detection algorithm than we would like. In the next few months, we plan on investigating methods for making predictions about expected behavior that are more robust or adaptive to these short-term effects.

Current work has focused on identifying univariate features that discriminate between normal and faulty behavior. However, the multiscale and multimodal nature of CENS deployments and the complex structure of the observed phenomenon suggests that the faults are likely to be much more complex than those that can be captured using univariate signatures. A next step will be to gain some intuition about when using multivariate information can improve detection power. Then the signature-based algorithm will need to be extended to include multivariate features, which will involve online estimation of a multivariate density as well as characterizing the expected dependencies or normal and faulty multivariate behavior.

## **SDP 09 Unblinking: Continuous Sensing and Its Implications for Modeling Uncertain Environmental Phenomena with Latent Geometric Structure**

### **SDP 09.1 People**

- Principal Investigator: Andrew Parker, Mark Hansen
- Faculty: Mark Hansen, UCLA, Statistics Department, Professor;
- Graduate Students: Andrew Parker, UCLA, Computer Science Department, Graduate Student Researcher

### **SDP 09.2 Overview**

In applications involving mobile sensors, we have often carried over many notions that are rooted in static sensing. The result is usually an application that uses the mobile sensor in a way that mimics subsampling a static deployment: the mobile sensor moves to a grid point where a static sensor might have been, dwells and then samples, moves to another point, etc.

An example of this is described by Singh et al. [1], where they mapped the water currents in a fresh water lake with a boat and an acoustic doppler. They partitioned the lake into a 256 by 256 lattice, and took samples at a subset of lattice points approximately 8 meters apart. If the sensor requires a long dwell time, sampling in this manner is entirely appropriate. Other times, the choice of sparse sampling is rooted in the classical power tradeoffs involving communication and computation found in static sensor applications, but these tradeoffs are not valid for platforms where mobility is the primary power consumer.

In Unblinking, we will consider the situation where we are free to sample (near) continuous paths using a single mobile node. We will focus on local sensors – as distinct from remote sensing – that can move within a 2-D plane. Finally, we will concern ourselves with environmental phenomena that are the result of processes that have an underlying geometric structure. In particular, we will focus on the problem of estimating the light intensity in the forest understory. There are at least two other current or recent projects at CENS that look at this same problem (Budzik et al., and Kong et al.), but they both rely upon multiscale sensing, where they use a camera to either generate tasks, or to suggest a partitioning of the field into subregions. The advantage there is that the camera offers a cheap, low fidelity global view of the phenomena, directing the more expensive, high fidelity PAR sensor to sample the field in some efficient manner. Our object is to explore methods where one cannot take advantage of multiscale sensing.

### **SDP 09.3 Approach**

We return to first principles by re-examining the nature of the phenomenon, and the capabilities of our sensor and mobile platform.

The light field can be modeled as the result of a three-step process. First, the underlying structure of the light field is due to the *geometric shadow* of the canopy, which is a projection of the canopy onto the forest floor as if the sun were a point-source of light. Second, the transition between the full sun regions, and the full shadow regions can be understood by considering the affect of the sun's finite angular radius and how its projection, when occluded by a straight edge (such as a leaf), results in a penumbral fringe. The intensity of light at a point in the penumbral fringe is a function of its distance between the boundaries of the full sun and full shadow regions. Finally, the third step accommodates spatial variation within the smooth regions of the light field. The variations account for light reflection from the sky dome, clouds, and nearby vegetation.

The light field is also characterized by its high temporal dynamics. A sunfleck can appear and disappear within a matter of minutes, and the predictable march of the sun across the sky results in somewhat predictable drift and skew of the light field apparent over longer observations.

Taking the three stage process and temporal dynamics into account, we propose a corresponding hierarchical model that accommodates each of these four components explicitly.

We also reconsider the capabilities of our sensor and mobile node. The LICOR LI-190 is the PAR sensor of choice for NIMS deployments. It has a fast response time and is typically sampled at 10 Hz and measurement noise is within 1% of the report value. Also, the NIMS-3D robot, with its cabling infrastructure for movement, enjoys accurate localization and positioning, with horizontal accuracy in the sub-centimeter range and location sampling at about 3 Hz. The combination of inexpensive and fast sampling, combined with accurate positioning allows us to contemplate near continuous sample *paths* as our sampling primitive. This represents a major departure from the usual point-based adaptive sampling schemes.

#### SDP 09.4 System(s) Description and Experiments

Our system is a simulation framework for driving a NIMS like robot around a 2-D field. The framework uses as input a stack of time-stamped images, such as the light field library of images previously collected by Diane Budzik, in order to allow testing with temporally dynamic phenomena. The framework keeps track of various physical variables to mimic the behavior of a NIMS node as it accepts commands to move and sample about a 2-D environment.

Using this framework, we've implemented two adaptive sampling algorithms published by other CENS researchers (Stratified Sampling, Batalin et al. & BioScope, Rahimi et al.) in order to compare new adaptive sampling strategies. Additionally, we have run many day's worth of experiments in order to understand the effectiveness of epoch-based adaptive sampling algorithms in the face of highly dynamic phenomena.

#### SDP 09.5 Accomplishments

Any epoch-based adaptive sampling algorithm will perform quite poorly when tested on the light field library. We proposed a degenerate epoch-based sampler that uses the ground truth image of the light field at the start of an epoch as the estimate for the state of the field the end of the epoch.

Perhaps multiscale methods can smooth out the variation in MSE of the resulting field estimates, but methods that rely on a single scale sensor needs to be adaptive in a continuous manner.

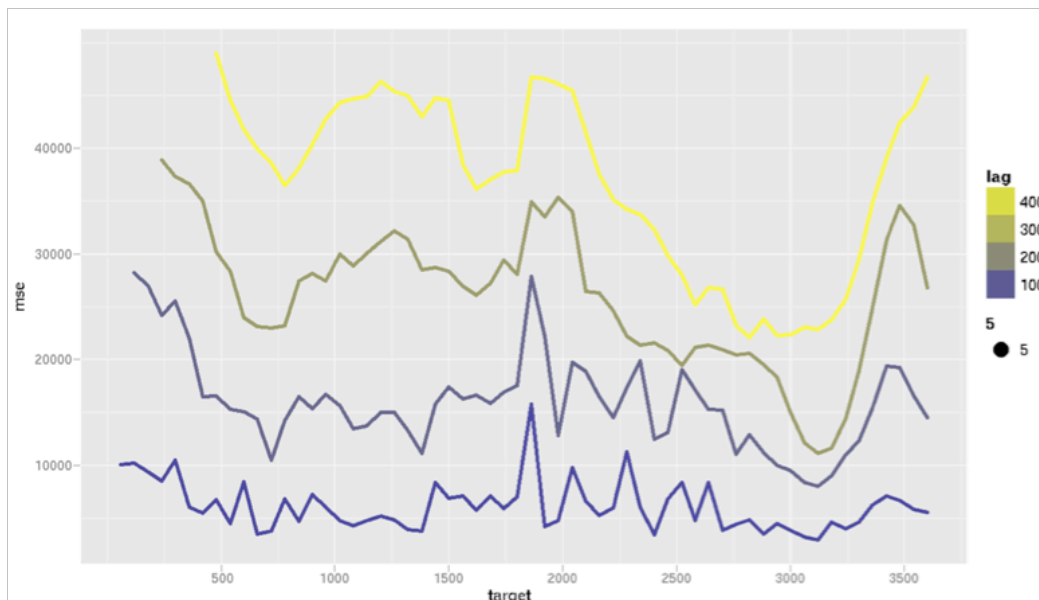


Figure 1: This shows the MSE over time of our degenerate epoch-based sampler. It demonstrates that any epoch-based sampler will experience wild fluctuations through out the day, and thus the expected performance of such samplers is unpredictable and highly dependent on the highly variable rate of temporal fluctuations.

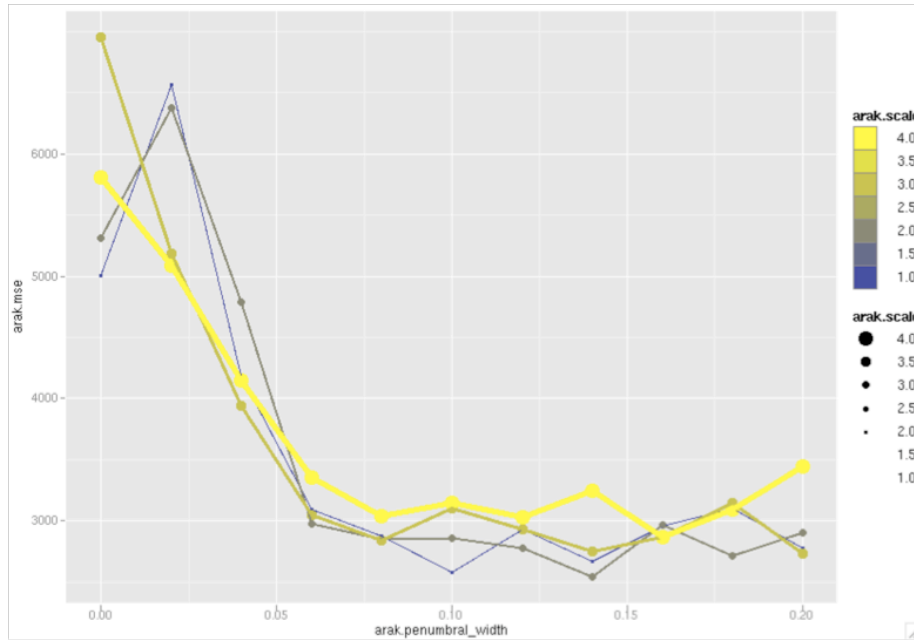


Figure 2: This demonstrates that effectiveness of adding a penumbral fringe to the basic PRF model. A fringe width of 0 is identical the original PRF model. Further experiments show that the MSE rises up again as the fringe width becomes much larger than the observations justify.

Our next accomplishment is the development and demonstration of polygonal random fields (PRF) as a viable spatial predictor for light fields. PRFs are especially competitive against predictors like bilinear interpolation when available sampled data is sparse. Further more, we have extended the PRF model to accommodate penumbral fringes. By this, we mean that while searching for a polygonal coloring that fits the data, the likelihood estimate that is calculated takes into account observations that may fall into penumbral regions. This extension of the model results in reduce MSE for its field estimates.

### SDP 09.6 Future Directions

In the coming year, will implement the next steps in our model and adaptive sampling scheme. The penumbral width is currently used as a global parameter, but naturally occurring light fields demonstrate variable width penumbral regions due to the variation in the height of occluding leaves and branches. The next step would be to allow for edge specific penumbral widths. A slightly less important, but still worthy task is to chase down the error apparent in the smooth subregions of the light field – the full sun and full shadow regions. Some multiscale task-based adaptive sampling algorithms completely ignore the shadow regions, and instead focus on reconstructing the lit areas of the field only. Other algorithms impose upper and lower bound thresholds on their light field estimates since plant photosynthesis has a non-linear relationship with light intensity, with virtual no activity below one threshold, and then fully saturated activity above another threshold.

The temporal aspects of the light field need to be modeled, and we believe there's promise in the vast body of literature surrounding cloud cover prognostication schemes, but this requires further investigation.

Finally, the field estimate will include an uncertainty field, over which we will define a utility measure for path samples over the field. We will then develop MCMC path generators to efficiently find paths with high utility value.

## SDP 10 Sensorbase: A Story in Three Parts

### SDP 10.1 People

- Principal Investigator: Mark Hansen
- Faculty: Mark Hansen (UCLA Statistics, CENS Area Lead - Statistics), Deborah Estrin (UCLA Computer Science, CENS Director)
- Staff: Richard Guy (UCLA Computer Science, CENS Staff)
- Graduate Students: Keith Mayoral (UCLA Computer Science, CENS GSR)

### SDP 10.2 Overview

Sensorbase provides a database backbone infrastructure for many of the sensing campaigns and projects that take part at CENS. It provides a mean to store all sensor data gathered in the field in a user friendly environment through the web interface, as well as providing an automated upload process using SOAP and RESTful services. With it, users can store any type of sensor data, be it text, audio, video or anything which can be represented in binary. To set itself apart from a simple relational database, Sensorbase also provides a user friendly web-based method of viewing and editing stored data, as well as a concept of a user's data. Keeping all data in a centralized location provides the benefit of allowing users to share pertinent datasets with other researchers. To ensure that only people who should see your data can, users can also specify the specific types of permissions to apply to each dataset or subset of. By providing a standard programmatic approach to retrieving data, users can process and manipulate their data straight from the Python or PHP scripts that use the data.

### SDP 10.3 System Description

Sensorbase, as mentioned before, is designed to be an online accessible system in which researchers can upload or *slog* sensor data to our central server running a MySQL instance.

The main interface to this system is through the website Sensorbase.org. The structure of the web interface is as follows:

**Users** must log in to the system through the front page of the site. No data or information related to any projects can be viewed by an anonymous visitor to the site. Each applying prospective user must submit an email declaring his or her area of research and how Sensorbase will be of use. Once reviewed, a user account is manually created for each applicant. This process ensures that only researchers have access to the data available through the website. Once a user has registered, she can start creating datasets for each project related to her research.

A **project** in Sensorbase is a collection of tables all related to a single real world research project. Within each project, individual **tables** are defined to further separate different datasets within a research project. These tables can contain any number of fields of varying types, from simple text and numbers, to images and sounds. This structure forms the basis of how datasets are stored and viewed in Sensorbase.

**Data Permissions** and privacy concerns are also managed through the website interface. There are four different types of permissions which can be granted either at the project level or at the finer table level. If a user has a greater permission granted to the project containing a table, then the project level permission supersedes the table level permission.

**Build access** gives a user read/write/create/destroy privileges on the specified project or table. Project creators start with build access, but can also grant it to other users or *collaborators*.

**Slog and read access** gives a user the read/write ability on the specified project or table. This is the equivalent of assigning someone as a *contributor* to your project without actually having control over the project configuration itself.

**Slog only access** is given to a user if you don't want them to have read privileges, but still require them to contribute data. This setting could apply in situations where you have users submitting personal information, like GPS location traces throughout the day, which would be considered private and should only be seen by the researcher(s) in charge of the project.

Finally, **Read only access** can be granted to individual users who you might want to share collected data with, but have no association with the project/table such that they don't need access to write or modify the data.

The permission levels described above define how specific portions of the data can be shared. Another privacy setting which applies to a whole project lets you declare whether the data collected is public or private. This affects whether or not an arbitrary user can search for and browse your data from the search functionality present in the website interface. If a user attempts to view your project when marked as private, they will be redirected to a page stating that access is not allowed. Projects marked as public automatically grant read access to all users of the website.

#### **SDP 10.4 Accomplishments**

Sensorbase has had many improvements over the few months which have helped it become even more useful as an online database.

**Integration of code** into the main Sensorbase build was done at the start of the year to incorporate different functionality which was developed by other CENS researchers in individual projects. John Hicks, a researcher at CENS added an Abode Flash based data graphing feature and a delayed query feature to allow for large dataset dumps to be scheduled to run at times when the site is not under heavy load. These features, as well as smaller cosmetic changes, were added to Sensorbase prior to the addition of SVN access described below.

**Programmatic control** of datasets has been one type of feature which was requested by Sensorbase users. While support for the programmatic retrieval of data already existed in Sensorbase, it became apparent that other features would be as helpful if there were implemented for programmatic access. These features, which were implemented as RESTful services, allow users to programmatically deal with projects and/or tables given that they already have proper permissions. Users can now create, update, delete, or modify the structure of a specified project or table. Furthermore, more advanced SQL queries are available through this programmatic interface, such as table joins. For a complete list of implemented features, one can visit the help page at <http://sensorbase.org/help/> as well as see example usage of each function.

**Availability of source code** is another aspect of Sensorbase which we've worked to simplify. We believe that the best way to have Sensorbase work for all users is to allow them to modify the base code to better suit their project's individual needs. If a feature which has been implemented in one of these separate projects can be generalized to work for other groups, then we can simply transfer the code over to the main distribution to make it available to all users.

In the past, when a group requested a copy of the Sensorbase source code, they would simply get an archive containing a prepackaged version of Sensorbase to set up on their own. We have recently set up an anonymous SVN repository which allows for anonymous read access of the source code for the current stable version of Sensorbase. When users download the source, they can find a text document to assist in the setup of their own Sensorbase build. By providing this resource, we can allow outside developers to configure Sensorbase for their own uses as well as contribute to the improvement of our main build.

**Improving database speed and storage** has been a large milestone for this year. We started the process of transferring over the underlying MySQL database and the vast directories containing all the binary data associated with Sensorbase to a much larger Sun Sunfire server named Thumper. This server currently has 24TB of space so sufficient space for data storage won't become an issue anytime soon. Nonetheless, we have gone about restructuring the directory structure of the folders which contain all the binary data such as images and sounds. This was due to the fact that on the old server, there was considerable delay added to file access because of the

number of files stored in each directory. By restructuring the database and separating the database backend from the UI frontend, we hope to increase the overall speed of queries and improve the performance of the site overall.

Many other smaller bugs and fixes have happened over the year as they have been pointed out by the users of Sensorbase which as a whole keep us just as busy as the larger issues. Users can always send requests for features or bug fixes to [sb-support@lecs.cs.ucla.edu](mailto:sb-support@lecs.cs.ucla.edu).

### **SDP 10.5 Future Directions**

Much of the work planned for the coming year is based on what features the research groups at CENS will be requiring for their respective projects. We have been in talks with groups from all parts of CENS to assess the requirements each group will need for their projects to interact with Sensorbase effectively. Some of the groups involved in talks include: CENS Deployment Center, Urban Sensing, NetworkedNaturalist, PeruNet, as well as others. The features they request will have highest priority in terms of new features to implement.

Adding a layer on top of Sensorbase which interacts with the database to notify users of configurable alerts is also a feature we will be implementing in the coming year. Allowing for an email to be sent to a user when a data entry passes a threshold or a project attribute changes would help users be better informed about the data entering their projects, especially when much of the data is automatically being slogged by the sensing devices themselves.

#### *Summary*

As stated in the sections above, many research groups both in CENS and out, use Sensorbase to store the data collected in their studies. Sensorbase plays a vital role in simplifying the process of organized data collection, processing, and dissemination. By continuously expanding upon the feature set and usability of the system, we can provide ongoing support for the sensing research community.

