

2.8 Statistics and Data Practices (SDP)

Statistics and data practices have transitioned from relatively “offline” to “critical real time” components of CENS systems and practices. Our developments in this past year are in the context of mobile and social computing systems and designing for data quality, use and reuse.

Mobile devices and social computing

Three projects employ mobile phone technology for wireless sensing applications in very different ways. *your.flowingdata* is a participatory sensing application that can be used for health monitoring and a variety of social activities (Figure 6). What’s Invasive is a participatory sensing application for citizen science that enables individuals to take geotagged images of invasive species and contribute them to a shared database for tracking and control of these plants. The Mobile Scientific Data Collection project is an application to assist scientific teams in collecting data in the field more cleanly, quickly, and simply than with traditional notebook devices. *your.flowingdata* is built directly on the Twitter platform, aggregating tweets from individual users into a database, and providing visualization tools. The Mobile Scientific Data Collection project also is based on micro-blogging, although not the Twitter platform, and enables scientists in the field to document their data and activities with short statements.

Capturing cleaner data

Several projects aim to improve the data capture and management process. The Anomaly Detection project enables scientists to identify patterns in sensing data in real time, thereby allowing them to modify their methods in the field and to correct for errors in faulty sensors. The CENS Deployment Center captures information about deployment activities, ranging from planning and equipment to documentation on the successes and failures of those activities. The social and technical knowledge gathered in the CENS DC provides context for data collected in the field, thus improving the reliability of current and future deployments. Similarly, the Mobile Scientific Data Collection project allows the real time capture of documentation on data and activities.



Figure 6. *your.flowingdata* website.

Data use and reuse

By capturing cleaner data earlier in the scientific and technical life cycle, CENS data become more usable by the teams that produced them and more amenable to reuse by others. The Monitoring, Modeling, and Memory project is studying the entire life cycle of CENS data, as part of a four-university partnership. CENS data activities are being compared to several other large, distributed, collaborative scientific projects. The results will provide guidance on best practices to each of the participating centers. CENS publications, along with posters, working papers, and other scholarly products, are being deposited or recorded in the eScholarship repository of the University of California. CENS is now one of the largest of the eScholarship sub-repositories. Our scholarly products are now exposed to the

larger online community, and readily discoverable by search engines. Upon the request of the 2009 NSF Site Visiting Team to expose our datasets as well, we launched the Data Discovery Library project. The Data Discovery Library is being constructed as a layer on the CENS Deployment Center, which also leverages that project. We streamlined the workflow of the CENS Annual Report process by requiring CENS participants to deposit publications and datasets, thus capturing them for reuse.

Data modeling

Modeling of sensing data, is critical at multiple scales. At the most micro scale, the Unblinking project models environmental phenomena with latent geometric structure. At an intermediate scale, the Data Conservancy project is modeling scholarly products of, by, and about the astronomy community in an effort to design systems for long-term data curation. The most encompassing data modeling project is the Object Reuse and Exchange effort, whose aim is to model relationships among all extant types of scholarly artifacts. The ORE project is part of a larger international standards activity, part of the Semantic Web, to aid in the linking and discovery of scholarly products. CENS is among the first to test the ORE on scientific data and publications.

In sum, the Statistics and Data Practices area of CENS has matured in productive and integrative ways to study phenomena throughout the scientific life cycle.

SPD 01 Anomaly Detection

SPD 01.1 Overview

Wireless sensor systems have significant potential for aiding scientific studies by instrumenting the real world and collecting measurements, with the aim of observing, detecting, and tracking scientific phenomena that were previously only partially observable or understood. However, one obstacle to achieving the full potential of such systems is the ability to process, in a timely and meaningful manner, the huge amounts of measurements they collect. Given such large volumes of collected measurements, one natural question might be: Can we devise an efficient automated approach to identifying the “interesting” parts of these data sets? We can view identification of such “interesting” or “abnormal” measurements (or events) in collected data as anomaly detection.

A good anomaly detection method should have the following properties. First, it should be able to accurately identify all types of anomalies as well as normal behavior (i.e., it should have low false negative and false positive rates). Second, it should be robust, i.e., the methodology should be relatively insensitive to parameter settings as well as pattern changes in the data sets. Third, it should require relatively small amounts of resources, as these are typically limited in sensor systems. That is, to run on sensor systems, it should ideally have low computational complexity, occupy little memory space, and require little transmission power. Last, it is also desirable for a detection algorithm to be able to detect anomalies in real-time or near real-time. This is particularly important for sensor systems corresponding to temporary deployments (as it might not be as useful to detect anomalies once the deployment is over) and those monitoring hazardous natural phenomena (e.g., spread of contaminants in aquatic ecosystems), where prompt detection (and reaction) can be essential to reducing loss of life and money.

We formulate the problem of anomaly detection in sensor systems as an instance of identifying unusual patterns in time series data problem. The basic idea behind our approach is to compare the collected measurements against a reference time series. We propose an anomaly detection algorithm, termed SSA (Segmented Sequence Analysis). We perform an extensive study using data sets from real deployments, which illustrates that our approach is accurate, robust, as well as efficient. We also show that our (online) SSA-based approach is more accurate than potential other (offline) techniques, which are more computationally intensive.

SPD 01.2 Approach

In our work, we are interested in the class of data collection sensor systems, where each mote (usually) collects periodic sensor data, possibly performs some local processing on the data, and then transfers the resulting data over multiple hops. We model the measurements collected by a sensor m as a time series $D_m[t]$, $t = 1, 2, \dots$

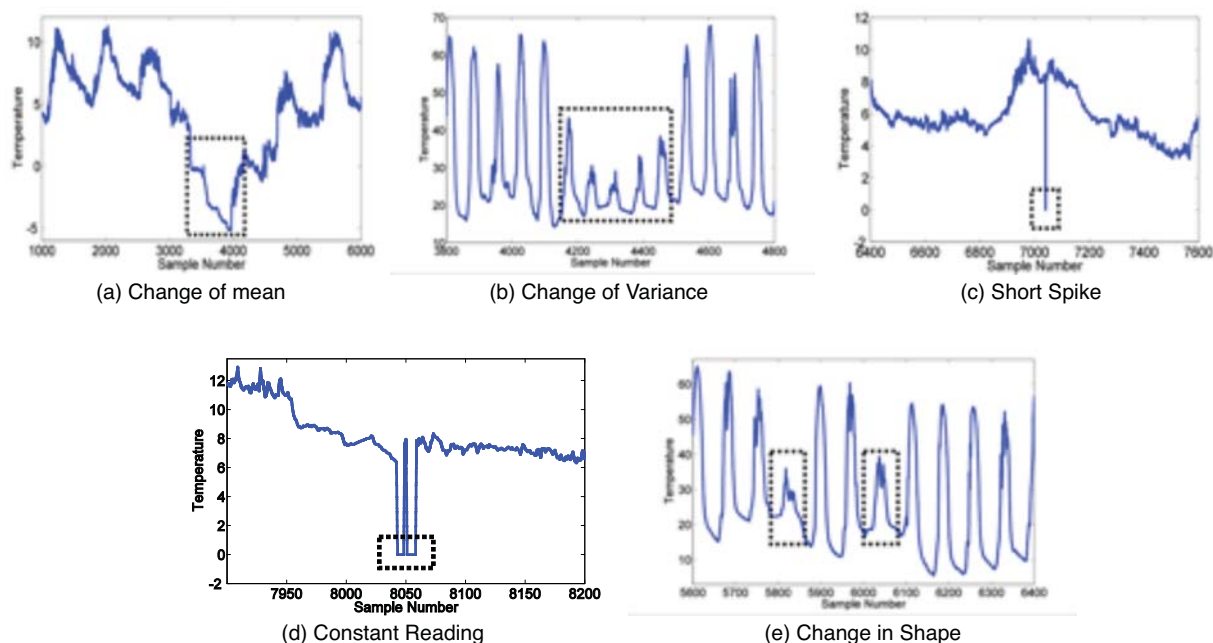


Figure 1. Categorization of anomalies

As already noted, we propose SSA as our approach for anomaly detection. We now briefly describe this algorithm.

At a high level, our approach tries to quantify how similar is a time series of sensor measurements to a given “reference” time series. Suppose we are given two time series, $D_{new}[t]$ and $D_{ref}[t]$, where $D_{new}[t]$ is the time series of new sensor data, and $D_{ref}[t]$ is the reference time series. SSA first constructs models corresponding to $D_{new}[t]$ and $D_{ref}[t]$. Then it compares these two models using a similarity measure. If the model for $D_{new}[t]$ is not sufficiently similar to the model for $D_{ref}[t]$, SSA concludes that there are anomalies in the time series $D_{new}[t]$.

We also note that SSA itself is not able to detect all types of anomalies; specifically, it is better suited for long-term anomalies. Thus, we propose a hybrid approach. The idea is combining SSA with Rule-based methods. The rules we use are based on earlier efforts, intended for fault detection in sensor data (where faults can be viewed as short-term anomalies).

SPD 01.3 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 real-world datasets. The sensor data time series used in our evaluations come from two sources: the SensorScope data sets, and the Jug Bay data sets from the Life Under Your Feet project. To obtain the ground truth, we visually inspected these sensor data time series, to identify both long and short duration anomalies. For ease of results presentation, we categorize the anomalies identified through our visual inspection into five groups. Figure 1 shows examples of these anomalies.

The accuracy results of our hybrid approach on all data sets are given in Tables 1 and 2. Note that in these tables, the x/y number indicates that x out of y anomalies were detected correctly (corresponding to $y - x$ false negatives) plus we also indicate the number of corresponding false positives (FP).

Tables 1 and 2 show that our hybrid method is able to detect both long duration and short duration anomalies, with a small number of false positives, and often without any false negatives. We would like to point out that long duration

Data Set	Change in Mean	Change in Var	Change in Shape	Short	Constant	FP
SensorScope 1	3/4	0/0	6/7	90/90	2/2	7
SensorScope 2	8/8	0/0	6/7	86/86	6/6	6
SensorScope 3	7/7	2/2	9/10	64/64	5/5	12
SensorScope 4	5/5	2/2	12/13	220/222	13/13	27
SensorScope 5	6/6	4/4	9/9	726/819	34/34	0
SensorScope 6	7/7	0/0	8/10	206/206	1/1	7
SensorScope 7	8/8	4/4	12/12	555/567	54/54	0
SensorScope 8	6/6	0/0	9/10	243/243	2/2	4
SensorScope 9	6/6	2/2	11/12	65/65	23/23	6
SensorScope 10	5/5	0/0	10/12	46/46	2/2	3
SensorScope 11	7/7	0/0	8/10	122/122	1/1	6
SensorScope 12	7/7	0/0	11/13	84/84	13/13	7
SensorScope 13	8/8	2/2	13/14	250/250	15/15	5
Total	83/84	16/16	126/139	2767/2864	171/171	100

Table 1. Hybrid Method: SensorScope

Data Set	Change in Mean	Change in Var	Change in Shape	Short	Constant	FP
Soil Moisture 1	7/8	1/1	8/10	53/55	1/1	0
Soil Moisture 2	8/9	1/1	9/11	74/74	1/1	2
Soil Moisture 3	5/5	0/0	6/7	42/42	0/0	4
Box Humidity 1	2/2	4/4	18/19	15/15	0/0	2
Box Humidity 2	5/5	7/7	27/28	16/16	2/2	2
Box Humidity 3	3/3	2/2	1/1	17/17	2/2	3
Total	30/32	15/15	69/76	217/219	6/6	13

Table 2. Hybrid Method: Jug Bay

anomalies such as change of mean and change of shape occur quite often in the SensorScope and the Jug Bay datasets. And, our approach is able to accurately detect these long duration anomalies.

We also performed experiments on our approach's sensitivity to parameters. We found that SSA is fairly insensitive to its parameter settings, given that these parameters are within a reasonable range. We are also able to show that our approach is robust to data faults, in a sense that it is not affected by short duration anomalies.

We are comparing our approach with other anomaly detection methods such as an ARIMA based approach, PCA based approach, and KNN (K-th Nearest Neighbor) based approach. Our initial results indicate that our approach outperforms all these methods in detecting long duration anomalies such as change of mean and change of shape anomalies.

SPD 01.4 Accomplishments

In summary, during this reporting period:

- We proposed an approach to anomaly detection in sensor systems that is able to detect anomalies accurately and in an online manner.
- We performed an extensive study using data sets from real deployments, which illustrates that our approach is accurate, robust, as well as efficient.
- We showed that our (online) SSA-based based approach is more accurate than potential other (offline) techniques, which are more computationally intensive.

SPD 01.5 Future Directions

We believe that our work opens up new research directions in automated high-confidence anomaly detection and classification. We plan to more extensively compare our anomaly detection framework with existing techniques (e.g., those used in network anomaly detection), as well as to implement our approach in the context of a sensor system architecture and extensively study its computational, memory, and communication requirements.

Moreover, we plan to extend our work in the following directions: (1) investigation of dynamic settings of parameters, (2) feedback mechanisms between the local and the aggregation steps of our approach, and (3) investigation of other techniques that are useful in our hybrid framework.

SDP 02 your.flowingdata

SDP 02.1 Overview

your.flowingdata (YFD) is an application that lets people collect data about themselves and their surroundings with the popular micro-blogging service, Twitter. Twitter asks the very simple question: "What are you doing right now?" The user updates friends, family, and others in her network on what they she is doing using succinct messages, or tweets, in 140 characters or less. YFD uses this short message paradigm with a simple syntax, which in turn allows users to explore their daily habits via an online visual interface. YFD has about 3,000 users who log computer activity, music-listening, eating habits, weight, blood sugar, and many others. It serves as both a way to track short-term goals and keep a long-term data journal by intertwining data collection with everyday activity.

SDP 02.2 Approach

YFD was designed to fit in with regular Twitter habits. Millions of people use Twitter to update their network on what they are doing or what is going on around them, so the online culture of posting smaller bits of the everyday already exists. With this in mind, we can think of YFD as an extension of Twitter. Users send messages to YFD via direct messages, which are private messages, on Twitter similar they would send a regular tweet.

The point was to make data collection as easy as possible so that users could focus on analysis with the visualization tools available on the YFD site (<http://your.flowingdata.com>). Interaction with the data was most important in design of the site. We wanted users to be able to explore their data beyond static staistical charts. Visualization started with familiar interfaces like a calendar and tag cloud. Users can then explore their data with searchable tools for a broader view of trends, patterns, and relationships. Finally, other tools were designed specifically to highlight certain aspects of user data.



Figure 1. your.flowingdata homepage, <http://your.flowingdata.com>

SDP 02.3 System Description

There are three main components to the YFD application: connection with Twitter, application framework, and the visualization toolset. The first component collects and parses messages from Twitter, the application framework provides user authentication and application and layout, and the visualization toolset, the focus of YFD, lets users explore their data.

Twitter

YFD has a special Twitter account that has the sole purpose to receive direct messages from users. Making use of the Twitter API, formatted messages are parsed and then stored in a MySQL database. The main challenge was creating a syntax that was flexible enough to allow different types of data collection, but at the same time was felt natural enough for users to pick up quickly.

All messages follow this format:

```
d yfd <action name> <value> <unit> (at) <timestamp>
```

Everything after <action name> is optional. An example will make this format clearer. To start simple, if a user were to track eating habits, she could use the action name of "ate." The message might look like this:

```
d yfd ate
```

The first part, "d yfd," tells Twitter to send a direct message to username "yfd." The second part, "ate" is the data. The user logs when she ate. If, however, the user wanted to track not just when she ate, but what she ate, she could include value and units.

```
d yfd ate 2 hot dogs
```

In the above example, again, "d yfd" is part of Twitter's syntax, "ate" is the <action_name>, "2" is the <value>, and "hot dogs" is the <unit>. The time of the action is automatically logged as the time of the tweet. To set the timestamp to something in the past, the user adds "at <timestamp>" to the end of her message. For example:

```
d yfd ate 2 hot dogs at 6:30pm
```

With this simple yet flexible syntax, the user can track a number of things. The user can also use hashtags (i.e. a word preceded by a '#') to categorize her data.

Application

Once the message is sent, it is parsed and stored. The user can then login to YFD to view her data. YFD uses the OAuth protocol for authentication, so that users can login with their Twitter username and password. The application itself was implemented in Django, the Python Web framework.

All data is kept private, unless it is made visible in the user-enabled public views. Initially, there were no public views, however, many users requested a way to share their data with others, so a way to piece together pre-selected modules for existing data was created. Some used the public views to update others on a status much in the same way people use Twitter, but from a data point of view. One user for example, updated his partner on when he was going to sleep and waking up, because he promised her he would maintain a more regular sleep schedule.

Visualization

YFD provides several interactive visualization tools that let users explore their data. As said earlier, users are first presented with a familiar interface in calendars and tag clouds, and then given the option to look at their data from other views. A searchable stacked area chart lets users investigate patterns and relationships over time. A treemap provides a view into aggregates, which like stacked area chart, is searchable and interactive.

Finally, there are two visualizations that display specific aspects of the data. The first shows the duration between

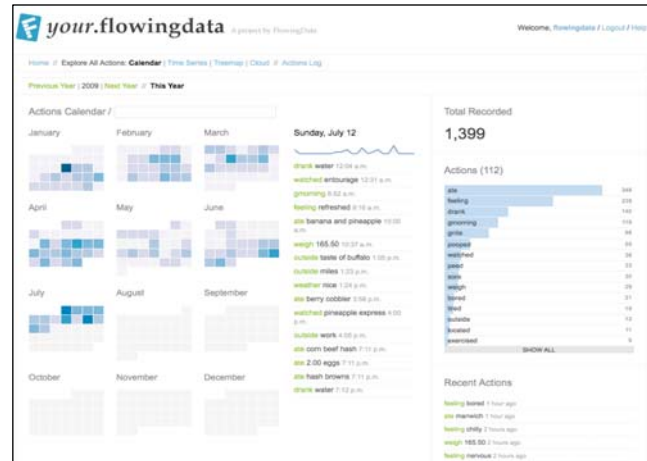


Figure 1. A calendar view provides users with a familiar interface to their data. Days are colored by number of times an action occurred. The calendar updates dynamically as queries are typed in a search box.

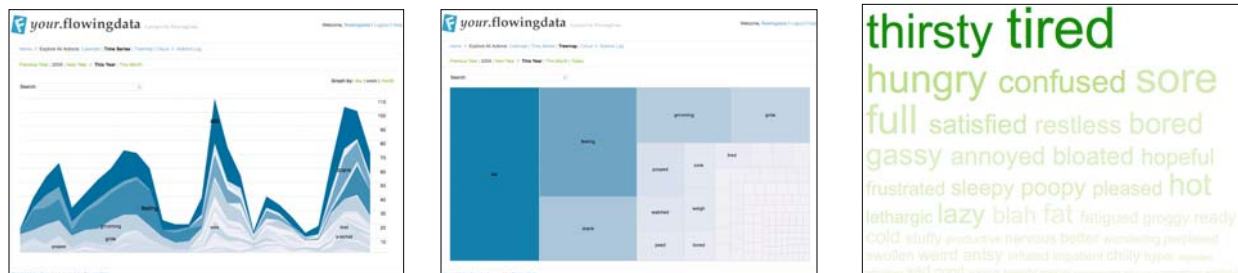


Figure 3. The stacked area chart (left) lets users explore patterns over time while the treemap (middle) and tag cloud (right) show aggregate values.

actions. For example, the user might track her sleep patterns by sending a message when she goes to sleep and when she wakes with “goodnight” and “gmorning,” respectively. The durations visualization shows the time in between the two actions, or more specifically, the amount of time the user sleeps each night.

The second visualization explores cross-correlation between actions. Many actions that users log often correlate. As a simple example, one user keeps track of when he wakes up and what he drinks. There is a strong correlation between wake up time and when the user has his morning coffee. Similarly, the user often has a beer at the end of the day, so there is bigger time gap between wake up time and when the user drinks a beer, but the relationship is highlighted in the graphic.

Again, the main purpose of these tools is to provide users with a way to explore and interact with their data, and in turn, make conclusions about their day-to-day activities.

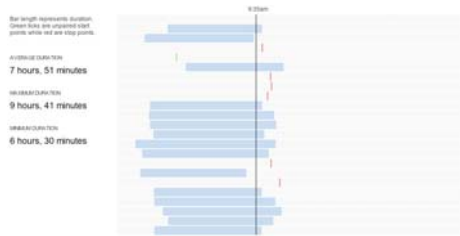


Figure 4. A durations visualization lets users explore the time in between actions.



Figure 5. Users can look at correlations between actions.

SDP 02.4 Accomplishments

The current version of YFD has been online and available for public use for seven months. Over 251,000 data points have been logged by 3,170 users. Many users found YFD through press from The New York Times and CNN. Others have found YFD through other Twitter users sharing with their network.

While most uses like health-related tracking were expected, there were also some unexpected uses of the service. Some made use of the Twitter API to automate data logging on YFD. Metrics included weather during a local heat wave, Web-browsing behavior and search, music-listening via the service last.fm, and computer hardware metrics such as temperature and download rate.

Twitter’s API flexibility has also allowed users to develop third-party applications to make data-logging easier on YFD. An iPhone application and Android application were developed independently along with ways to enter data with Mozilla Ubiquity and bookmarklets in Web browsers.

SDP 02.5 Future Directions

There are plans to provide a deeper visualization toolkit as well as expand the data entry syntax. YFD also provides reminders and notifications by sending direct messages back to users; however, there is still a lot of potential for “smarter” reminders. As it is now, YFD is more of a one-way application that users can send data to (with the occasional reminder). The hope is that we can expand on this to provide more feedback to users about the data they are logging and make YFD a two-way application. Ultimately we want to integrate data collection, interaction, and analysis into the everyday.

SDP 03 What's Invasive!: A Model for Spatializing Data Sources

SPD 03.1 Overview

"What's Invasive!" can be viewed more broadly as test case of the OurPixel ideals put into motion. OurPixel's goals are to promote citizen science via easy-to-use data sources and tools that traditionally have not been easily accessible to the general public. "What's Invasive!" is focused on leveraging geospatial tools and smart-phone capabilities to help communities anywhere help locate invasive plant species. By making geo-tagged observations through these devices, users can not only track them via the web, but also interact with the data. For this reason, it was necessary to build a standard set of geospatial and visualization tools that could be used in any number of projects.

This discussion will outline the design of a highly flexible open-source web-based mapping system that is to meet the data demands of citizen-science campaigns. With the recent developments in open-source software for geospatial data, we have developed a system that is easily expandable, enables users to access GIS data layers and interact with submitted data.

SPD 03.2 Approach

The breadth of a system that embodies OurPixel required considerable input from team members. Several rounds of proposals went out in order to encapsulate an open review and acquire group input. It was important that our choices were dictated by ease-of-use, compatibility between various software components, and tied together by a unifying framework. Some of the requirements for such a system are as follows.

We want tools that are open-source, allow for rapid development, and interoperability. These should not be unique to "What's Invasive!", so what is built is easily replicated for other projects. This system should be an expandable framework in a way that accommodates changes in a simple, efficient way.

Because of the geospatial nature of "What's Invasive!" (and more broadly to OurPixel), it is desirable to have intuitive and unrestrictive geospatial tools that allow for easy access to data layers such as shapefiles, raster data, and KML. It is important that users can create and store data, tag points, draw polygons, upload photos, through an effective user-interface. For data storage, we need a database that should be "spatial", so that we can take full advantage of the location-based nature of the data.

Social-networking features are highly desirable and should reflect the goals of the OurPixel project. This includes an ability to create a community around the project. An implementation might include tools such as blogs, photo management, groups, and discussion boards.

In terms of system design, Figure 1, these features fall into 4 categories, namely,

- Server-side software and Databases
- Middleware and Web Framework s
- Geo-spatial tools and plugins
- User-Interface

SPD 03.3 System Description

Back-end and Database

The back-end of the system is built on top of a Ubuntu 9.10 Linux-Apache-M. This allows for easy access to our required open-source software and is a very stable platform. Data is stored through the use of PostgreSQL and PostGIS. PostGIS provides an interface for PostgreSQL to spatialize our data. It is highly interoperable with our other choices, in order to accommodate a natural flow between components.

Web Framework

The key component that connects all of our tools is Django. It is a high-level Python web framework that embodies rapid-development, functionality between components, and is open-source. It utilizes functional and inter-operable

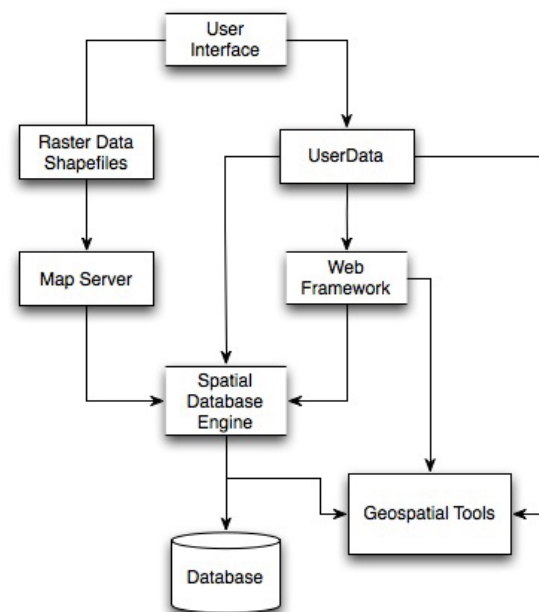


Figure 1. Base System design

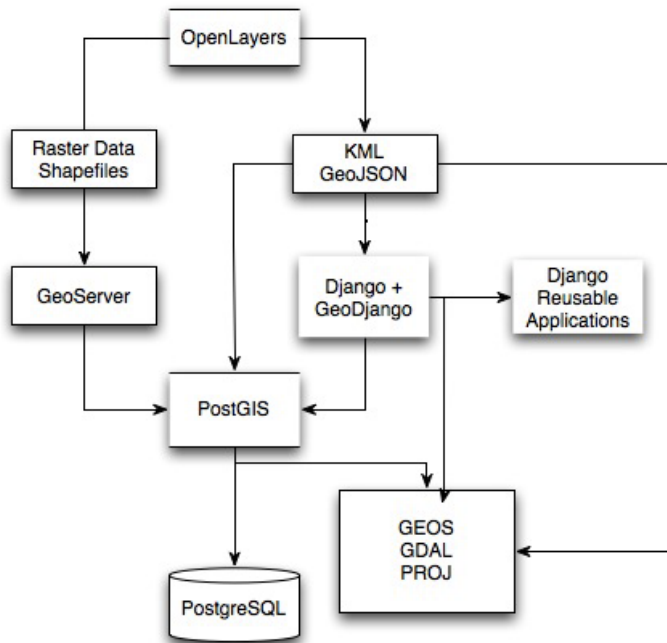


Figure 2. Implemented design

applications to speed up the development process, such as GIS and Social 2.0 features. It is run on an Apache web-server using the Web-Service Gateway Interface.

Geospatial tools

The GIS branch of Django, GeoDjango, provides a geographic web-framework for building GIS web-applications. It utilizes powerful geospatial libraries (GEOS, PROJ, GDAL) that work together with all levels of the system. GeoServer gives us a natural tool to serve up more complex data.

User-Interface

The OpenLayers Javascript library offers extensive support, extensibility and interoperability between our components (GeoDjango, PostGIS). It allows us to access any number of map tile servers such as Google Maps, Yahoo! Maps, and OpenStreetMaps. It has strong functionality that connects to all of our potential data sources.

Figure 2 gives an overview of how these pieces interact with each other.

SPD 03.4 Accomplishments

The core framework for an expandable geospatial-social system is complete. By using the Django web framework, we have built a core foundation that is easily modified and replicated for other campaigns. It is the heart of the system and connects to all other pieces in a simple, yet powerful way. It connects easily to our robust geospatial tools and opens up many avenues of future development. A map-server in the form of GeoServer is available to serve up shapefiles, KML, and raster data. Our user-interface utilizes OpenLayers as an effective way of connecting the user and data available. However, there are other components that still need further development; details for improvements on the current system are presented below.

SPD 03.5 Future Directions

There are several areas of development that would need to be implemented in order to fully expose the capabilities of the system and OurPixel's broader goals.

Social Networking Features

Due to time constraints, the social networking and community aspects were not implemented. This could be accomplished without many delays, because of the nature of the Django framework and Pinax. Pinax is a application plug-in for Django that offers a plethora of features to "socialize" any project.

Mapping-Geospatial Features

The current map interface through OpenLayers/Google Maps could be improved through the use of JavaScript frameworks, ExtJS or jQuery. Additionally, more data layers and a more intuitive interface to access these layers would be ideal.

Data Feeds

There should be better communication or integration between the current setup of databases and data sources. There is data in multiple locations, e.g. Twitter and Flickr, and connecting them would be parsimonious. This could be done via a reusable application with Django.

Analysis and Visualization tools

There is a need for advanced analysis and visualization tools for "What's Invasive!". These should go beyond basic charting and help user's explore the data in an intuitive and statistical way. The use of R via Python, in conjunction with Django, should provide a direct path for this.

SDP 04 Unblinking: Continuous Sensing and Its Implications for Modeling Uncertain Environmental Phenomena with Latent Geometric Structure

SDP 04.1 Overview

Related work for adaptive sampling mobile platforms (such as NIMS) have typically concentrated on point samples, often with the goal of minimizing the number of points sampled. The Unblinking project is differentiated by considering (near) continuous path observations. We have formulated an approach that we call Active Paths for generating multi-segment linear paths (MSLPs). It makes novel use of stochastic geometry machinery to propose length efficient paths that accommodate a priori knowledge of desirable path configurations. Active Paths is based on simple energy functions and a reversible jump Markov chain Monte Carlo (RJMCMC) sampler, and shows promise when tested against a deterministic sampling design on 2D boolean models, which has been used to model a variety of environmental phenomena.

SDP 04.2 Approach

Active Paths iteratively produces length efficient MSLPs given the current field estimate and past path observations. Generally speaking, the algorithm iterates between sampling the field, and updating the field estimate. An example of the evolution of the path over several iterations is shown in Figure 1. To do this, we define a density over the space of MSLPs that appeals to intuitive notions of what these paths should look like. We consider paths described as an ordered series of line segments on the 2D plane, restricted to the area under observation. We define the multi-segment path process in terms of energy functionals.

The energy functionals balance a trade-off between proposing paths that chase down uncertainties in the field estimate, versus achieving good spatial coverage. Uncertainties in the field estimate are reduced by considering the variability of supporting observations weighted by their distance from the location at which you wish to estimate the field. Spatial coverage is rewarded by defining a repulsive energy for overlapping segments of proposed MSLPs. And since we want to prefer simpler and shorter MSLPs when possible, we penalize paths as a function of length and the number of segments.

We will assume that the spatial phenomena under consideration is a realization of a Boolean model, which is the union of independent random compact sets (grains) placed at points (germs) corresponding to a Poisson point process. One of the most popular Boolean models is the Bombing model, where the grains are discs of random sizes.

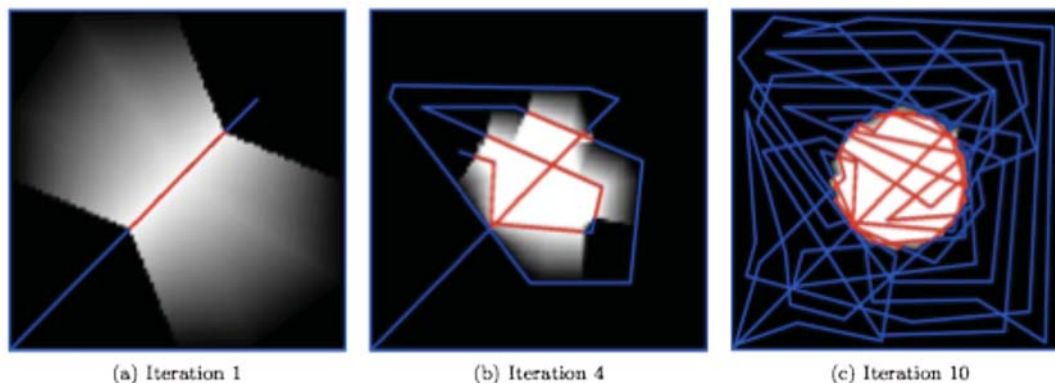


Figure 1. Above is depicted an example of the evolution of the path selected by the Active Path algorithm over several iterations. The path is superimposed over the bilinear interpolation of the observations.

Once the path observation has been made, we need to create an estimate of the underlying field. We have had favorable experience with bilinear interpolation, a point-based multivariate spatial prediction algorithm. It scales suitably with the amount and nature of the data we collect, and it makes few assumptions about the underlying phenomenon. The only draw back is that bilinear interpolation is point-based, forcing us to convert our path sample into a large number of collinear point observations. This may seem unsatisfactory given our desire to operate over paths rather than points, but the overall design of the Active Paths is independent of how a field estimate is produced.

We have also studied how the energy functional can be modified if our estimate of the underlying fields is a function of a sampler over the space of possible field configurations conditioned on data, such as for a bombing model. In this case, we have access to the Markov chain of field configurations. We have formulated two energy terms for the

bombing model in this scenario: one that rewards MSLPs that run parallel to the direction of greatest variation for individual discs in the Markov chain; and another that rewards MSLPs that are expected to traverse disc boundaries at (near) perpendicular angles.

SDP 04.3 System(s) Description and/or Experiments

We have implemented a reversible jump Markov chain Monte Carlo (RJMCMC) sampler of MSLP configurations that converges to the target distribution dictated by the energy functional for paths defined above. Our Active Paths implementation uses the sampler to drive the spatio-temporal NIMS simulator we developed in the previous year (though we do not take advantage of the temporal capabilities of the simulator in this case). We have run the Active Paths algorithm over images of sunflecks captured by Budzik et al., as well as synthetic images representing realizations of a bombing model.

SDP 04.4 Accomplishments

In our experience, Active Paths almost always performed measurably better than the deterministic raster scan sampling design. We compared our algorithm against raster scans over realizations of the bombing model, and the results for two different bombing configurations are shown in Figure 2. We used root mean squared error (RMSE) of the interpolated images against the true image as our measure of comparison. The relative performance depicted in these plots are typical of the other experiments we conducted on similar bombing model realizations. In particular, notice that Active Paths tend to seek out and refine boundary regions, while at the same time maintain uniform spatial coverage throughout the field.

Our algorithm, implemented in R, was able to achieve favorable results compared to raster scan even in terms of CPU time. This is an important consideration, because MCMC based techniques are notoriously slow, with run times often measured in hours. The algorithm was stopped after a total of 12 iterations, which resulted in a total path length of usually over 25 units, where the window of observation K was the unit square. The simulated annealing schedule for each iteration of Active Paths was 20,000 MCMC steps, but in our experience, 5,000 MCMC steps were often enough for the sampler to converge onto proposal paths with minimal energy, which required about half a minute of

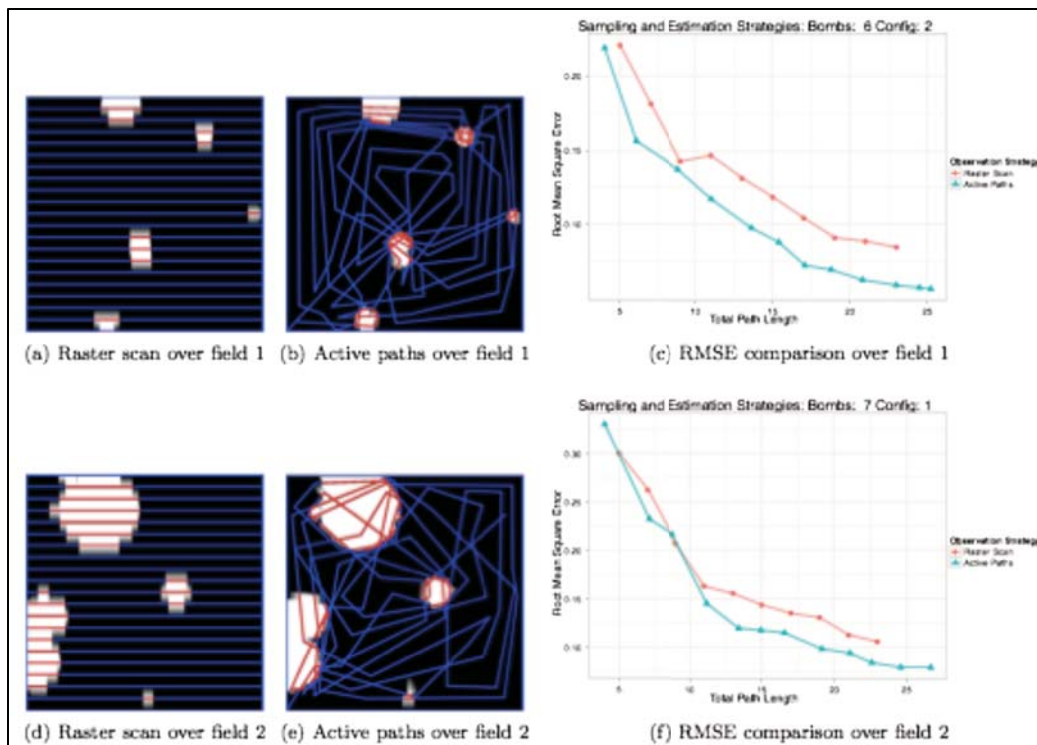


Figure 2. Comparison of raster scan versus Active Paths on two different boolean fields. The blue and red colorings on the line segments represent the values “observed” along those segments, with blue corresponding to 0, and red corresponding to 1, and is superimposed over the final interpolated field.

computation per iteration. And as the iterations accumulated, the sampler did not experience any slow down. In part, the computational efficiency was due to defining path energy functions that depend on previous paths only through potential fields that could be updated incrementally with each iteration.

We have recently submitted a description of Active Paths and the above results to a conference.

SDP 04.5 Future Directions

For future work, we will further develop techniques that incorporate knowledge of field models into the path energy functions. For example, if we know that the underlying field is approximately a union of discs, this imposes restrictions on the shape of boundaries, which we can use to better identify areas of uncertainty in the field estimate. We have already implemented a sampler for bombing model configurations conditioned on MSLP observations.

Also, we have noticed that the current algorithm selects paths that narrow down the location of discontinuities in the 2-D field in a way similar to binary search on a 1D step function, and it would be interesting to formalize this sequential search-like behavior. Finally, there is an opportunity to consider paths other than line segments, which may better reflect the natural paths drawn by different mobile platforms.

SDP 05 CENS Deployment Center and Data Discovery Library

SDP 05.1 Overview

CENS researchers develop flexible wireless sensing technologies that can be used in a variety of scientific applications. These technologies are used to produce valuable scientific data during real-world deployments. 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 and data collection best practices is a valuable asset for CENS. Our central research questions in the CENS Deployment Center project 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 focused our efforts this year on the second question, namely how to use our approach to make CENS data more discoverable and available for secondary use by potential outside users. The NSF is pushing grant recipients, including CENS, to make data management and data sharing a higher priority. In response to this push, we have designed a metadata repository that enables CENS data to be discovered by potential users. This metadata repository draws on our research and experience with constructing digital libraries for CENS information, and is being implemented as part of the new CENS annual report generating system.

SDP 05.2 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.

A parallel goal of our work is to add value to CENS data through descriptive and contextual information that surround the data collection processes. The data from CENS deployments are spatially and temporally irreproducible, having scientific value both to immediate research questions and long-term longitudinal studies. The CENSDC was our first attempt at devising a system to collect descriptive and contextual information about CENS data. Our work on the CENSDC informed our efforts this year in developing a repository of metadata about CENS data to make CENS data more discoverable by outside individuals. We focus on making CENS data more “discoverable” for a couple of reasons. First, CENS researchers collect a diversity of scholarly products that might be considered “data”, including images, audio files, physical samples, and numeric data in both digital and analog form. Second, these resources are distributed around community, lab, and individual computer systems. Some CENS research data and documentation are available online through lab websites, but many are in protected computer systems, personal laptops, or in file cabinets or refrigerators. Collecting and integrating all of the Center’s data into a single system would be prohibitively expensive and time consuming. Our “discoverability” approach offers a middle ground, collecting brief descriptions of data sets for the purposes of making them more visible and available to outside communities, without the costly overhead required to collect the actual data themselves.

This metadata repository, which we’ve titled Data Discovery Library, is expected to serve internal needs of CENS as well. The system will assist CENS researchers in keeping track of their own data resources and will provide the Center’s administrators with documentation of our research output. We also hope that it promotes the sustainability of CENS data and other scholarly products beyond CENS’ NSF funding. Thus, we are focusing on designing the system to be lightweight and easy to use with minimal assistance. Additionally, we are using open source software tools for the back end database and web display.

SDP 05.3 System(s) Description and/or Experiments

This section describes our system development work for both the CENSDC and the Data Discovery Library. System development for the CENSDC in the past year has primarily focused on collecting deployment data and investigating how to best implement interconnections with SensorBase. CENSDC data collection has focused on the CENS Seismic project in Peru, and the CENS projects at the San Jacinto Mountains James Reserve, and has involved talking with participants and following project email lists. Our work in designing interconnections between the multiple CENS information systems has included talking with the SensorBase administrators to see how the CENSDC and SensorBase approaches can fit together. Some results of this have been the addition of some CENSDC fields in SensorBase, such as date information, and development on modules that take advantage of the SensorBase API to allow CENSDC users to interact with SensorBase data remotely. This work is ongoing, and is described more in the future work section. For the CENS Data Discovery Library, our focus has been on developing a flexible set of metadata fields that can be used to describe CENS data. The fields in our prototype metadata schema for this system

were initially drawn from our database structure in the CENSDC. As we developed these fields, we found that our approach dovetailed with the Dublin Core metadata set (<http://dublincore.org/documents/dcmi-terms/>).

The Dublin Core metadata set was designed to be a simple and flexible descriptive schema for the discovery of digital resources, and thus is a good model for our approach, as shown in Table 1. We investigated discipline-specific metadata schemas for this project, such as the Ecological Markup Language and SensorML, but these are not used by CENS researchers in their current research, and are not flexible enough to serve the diversity of research and data types found in the center.

We are leveraging the existing NSF reporting cycle to acquire the data. As part of the annual reporting process, CENS' staff solicits information about a variety of scholarly products, such as publications, presentations, and new grants. This information is useful not only for the NSF, but as a means for CENS to document its return on investment to our many other partners and funders. This year each research group will also report their datasets using a new automated annual report system.

SDP 05.4 Accomplishments

This section describes our lessons learned in implementing the CENSDC, and the results of the pilot test of the metadata fields for the metadata repository.

CENSDC Lessons Learned

The process of implementing the CENSDC informed our data practices research and the design of the metadata repository in several respects. First, in implementing the CENSDC, we found that the culture of ad hoc planning is very strong in some of the field-based sciences, particularly the environmental and ecological sciences, which were our initial target users. Researchers in these disciplines conduct their field activities in a flexible way, adjusting activities to unpredictable weather, varied flora and fauna, and unreliable equipment. Thus, spending time creating a formal plan was not a high priority for some of our targeted users. However, most teams appreciated the assistance of a graduate student from the data practices team to work with them in the field and to help document their processes.

A second lesson was that deployment planning was conducted with a variety of tools that are not easily integrated. Email remains the main planning tool for most teams, with plans circulating via mailing lists. (CENS has about 100 dedicated distribution lists and about a third of them are used for deployment planning). Planning also relies on basic computer applications such as Word documents and Excel spreadsheets. Some research groups in CENS now use collaborative web tools, such as Google Docs and Spreadsheets, to track equipment lists and other deployment information. CENSDC development predated the wide availability of such shared online tools. While these tools serve some of the CENSDC functions, they are team-specific. An important goal of CENSDC is to share knowledge across the CENS community, as much duplication exists in the use of sensor technologies, other kinds of equipment, and field deployment activities across research teams and disciplines.

A third lesson in data practices from CENSDC is the diversity of deployment practices. CENS sensor deployments vary in duration, spatial extent, equipment complexity, number of people involved, focus, and outcome. Some deployments involved installing sensors for months at a time in static locations, and some deployments involved campaigns of three to five days with mobile sensor systems. Some research groups repeated the same kinds of deployments at multiple sites or at the same site, while other groups never perform the same activities on consecutive deployments.

Data Discovery Library Pilot Test

We performed a pilot test of metadata shown in Table 1 with four CENS researchers: two computer scientists, an engineer, and a domain scientist. The domain scientist and one of the computer scientists are part of the same research team. The participants in this test were chosen through targeted sampling of individuals who were known to have participated in original data collection, and as well as to sample from multiple disciplines and projects within the center. We asked these researchers to create metadata for the main data that they were using in their primary day-to-day research. We used a "talk-aloud" protocol, asking the testers to describe what they were thinking and writing as they completed the metadata descriptions.

The preliminary findings of the pilot test identify a number of issues that complicate the metadata creation task. Due to the limited scope of this pilot test, these are not meant to be definitive results; rather, they outline important issues that we will use as points for further investigation as the system matures.

- *Item in hand vs. distributed objects:* Much of CENS' data are not individual self-contained items. They may have many constitutive pieces, such as multiple files and database tables, and they may be spread around multiple locations, such as lab servers and personal computers. In creating metadata, researchers have to decide what is to be described as part of a single project or data set, and where to draw boundaries between data sets.

- *Non-self-describing resources*: Much of the data collected by CENS researchers are not textual, thus researchers must either create textual descriptions from scratch to describe image, audio, or numeric data, or they else adapt existing text from research publications or technical reports to the data description task.

- *Sense making*: Metadata fields may not make sense to a researcher who has not seen them before. In all four pilot tests, the researchers asked for clarification of what they were expected to include in particular fields, requesting examples or further explanation. Some fields, such as “permissions”, were problematic to all testers, while other fields, such as “size and format”, were only confusing to individual testers.

- *Projected/reverse sense making*: The potential users and uses of research data are often not obvious, even to the researchers who collected them. Researchers must project how their data might be used, and create metadata appropriate for those uses.

- *Talking vs. writing*: In describing their approach to filling out a specific field, particularly fields that they were less sure about, the pilot testers would “talk through” a field until they were surer about what to include. These verbal discussions about what should or should not go in a given field were not always reflected in what was written down. Often a rich verbal discussion resulted in a brief written statement.

- *Individual knowledge vs. group knowledge*: CENS research takes place in group settings. Individual researchers may not know what to include in certain fields, but do know who in the group to ask. For example, a couple of the pilot testers said that they would need to ask their principal investigator about how to fill out the “funding” and “permissions” fields. Another related issue is that different individuals in the same project may have different perspectives on what the boundaries of the data set are, and what descriptive information should be included. For example, the domain scientist and the computer scientists who are part of the same research team emphasized different parts of the same data. As part of the data description, the domain scientist emphasized the physical work involved in installing research equipment in the field and did not provide many technical details. In contrast, the computer scientist emphasized technical features of the data and the way it was collected, and gave no reference to the field work.

- *State of a project*: Different CENS projects are in different states of completion. Metadata has different importance at different stages of a project. At early stages of a project, creating metadata descriptions might not be very useful, due to low data volume or quality.

- Reflecting back on our initial goals—to make CENS data more discoverable, to help research groups keep track of their own data, and to develop a sustainable system – a couple of key challenges require further study. First, many of the issues identified above illustrate the lack of expertise that data authors have in metadata creation. This points to the development of training programs and more explanatory metadata creation systems, including examples and fuller descriptions of the metadata fields. Second, the ambiguity of boundaries around data sets and the fluidity of prospective users and uses of data suggest that training material and activities will need guidelines regarding the focus of the metadata creation process. Third, the tensions between individual and group knowledge suggest that we investigate metadata creation methods that include both individual and group

Data Description fields	Dublin Core element
1. CENS project name	title
2. CENS research group	publisher
3. dates (of collection)	date
4. place	coverage
5. people	-
contact person	creator
other participating researchers	contributor
6. data type	type
7. data description	-
research question (why collected)	description
what collected (variables)	description
data collection process and equipment	description
size, format	format
8. related publications (eScholarship URL)	relation
9. related deployment info (CENSDC URL)	relation
10. keywords	subject
11. location of the data (URL)	identifier
12. permissions	rights
13. funding source	source

Table 1. Metadata Fields used for Prototype Testing

contributions. And fourth, the varied states of project maturity suggest that we investigate the ways that metadata are, or can be, created piece-by-piece during the lifetime of a project.

SDP 05.5 Future Directions

Our work on the CENSDC and the metadata repository are part of our continued effort to understand how data collected in a collaborative research environment are created, managed, and shared. Our immediate work on these projects will follow two main thrusts. First, we continue to work on synergies to be gained by integrating the CENSDC and SensorBase systems. SensorBase can capture the sensor data per se, while CENSDC can capture the description of the deployment where the data were collected. If these objects can be linked effectively, each becomes more valuable. Further, the data and deployment information can be linked to resulting publications, embodying the value chain of those scholarly activities. We have developed the conceptual framework for instantiating these relationships using the Open Archives Initiative Object Reuse and Exchange protocol, and are working on the technical implementation.

Second, testing and implementation of the metadata repository will begin in late February, 2010. Evaluation of the system performance and the metadata records created by CENS researchers will follow. This evaluation will inform the next iteration of the annual report system as a whole, and the Data metadata module in specific. Over the longer term, we plan to extend this study of metadata creation by data authors in a number of ways. We plan to perform targeted interviews with data creators focusing on understanding their current metadata practices in their own work. We will ask researchers what “metadata” means to them, what their current data description practices are, and what is involved in sharing their data with people both inside and outside their research group (including the role of metadata in that process). As part of this, we will ask to see the data organization schemes (folder structures, naming conventions, database layouts, etc) currently used by research teams and perform content analysis on these schemes. This will help to characterize the typical state of personal data archives in distributed research environments.

SDP 06 Monitoring, Modeling, & Memory: Dynamics of Data and Knowledge in Scientific Cyberinfrastructure

SDP 06.1 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 06.2 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 06.3 System(s) Description and/or Experiments

To answer these questions, we are in year two of 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.

At the end of summer we wrapped up our first year with a research retreat at UCLA, bringing together the faculty and graduate student researchers for 3 days to discuss data collection, new methods, and how to coordinate our data analysis and writing projects across the sites for the coming year. This is the first retreat of three to support an iterative cycle of comparison of the cyberinfrastructure projects. More interviews were collected and transcribed over the months that followed. And by this summer we will have presented and published some preliminary results.

SDP 06.4 Accomplishments

- Collected first round of interviews from 20 participants
- Processed interview transcripts
- Developed multiple iterations of collaborative coding methods
- Developed multiple iterations of codes (question-based, core, expanded core, local/global)
- Coded transcripts for a shared database of coded data
- Ran a 3-day research workshop where researchers were able to experience CENS as a research site
- Borgman and Wallis launched new graduate course on "Data, Data Practices, and Data Curation."

SDP 06.5 Future Directions

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

- Develop a new interview instrument
- Conduct another round of interviews
- Perform participant observation
- Collect and review publications from CENS researchers
- Map and compare biographical trajectories of 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

This project is funded by NSF award 0827322 (start date 10/1/08; End date 9/30/11), 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)

SDP 07 eScholarship Repository

SDP 07.1 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 07.2 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 07.3 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. The repository also serves as a platform for generation social network analysis data.

SDP 07.4 Accomplishments

- 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
- To support CENS authors who need to upload their works, we have developed an Annual Report Upload System that supports all of the data gathering activities that are part of the Annual Report writing process, including a part to capture the bibliographic data necessary for depositing new publications in eScholarship

SDP 07.5 Future Directions

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

SDP 08 Mobile Scientific Data Collection

SDP 08.1 Overview

When building data digital libraries, understanding the context in which data were collected is critical to understanding and using the resulting data. Contextual data are essentially "metadata" that describe the data themselves. The challenge of capturing and collecting contextual information in dedicated digital libraries is compounded by the various and amorphous conceptions of "context" itself. Discussions of context are better conceived as discussions of "practices". Contextual information about a scientific observation or experiment is, indeed, a description of scientific practices: for example, the data collection process, the equipment used, researchers involved in the observation, and the exact location of an observation. These types of contextual data are often only minimally described or left out entirely from the presentation of research results in scholarly publications, which are often all that researchers have when finding and using data collected by someone else. Researchers apply their own field expertise to evaluate data collected by others for relevancy and identify any potential problems. Data digital libraries will certainly help with accessing data, but as long as researchers must rely on often incomplete published reports of the data collection processes as data quality verification, data digital libraries will not meet their promise.

In this report, we discuss our ongoing work in addressing this challenge in the context of field-based research that use Wireless Sensing Systems (WSS). WSS can produce data at greater volumes and higher resolutions than ever before, allowing researchers to observe previously unobservable phenomena. With the addition of WSS, field research in areas such as environmental biology, seismology and urban sensing is becoming highly instrumented, computational, and collaborative. The utility of data collected in situ via wireless sensing systems increases when coupled with contextual data that describe the setting in which the observations were taken: the equipment used to collect data, the researchers involved, and specific environmental conditions. Ensuring capture, description and preservation of these data is a fundamental task for scientific information management, as large volumes of sensor-collected data are linked to specific times and places and are thus irreplaceable. For these data, losing descriptive contextual information means losing the data altogether.

Our proposed solution is to capture contextual information by providing researchers with tools to document and adjust their research methods and data collection practices as they interact with each other, the experiment location, the data collection equipment, and other constraints such as time and money. Because scientific practices change significantly from domain to domain, providing researchers with reliable tools and techniques to describe their data collection practices can be difficult. This is especially true for field-based sciences because of the inherent variability of real-world locations. Contextual data might vary greatly depending on the type of data collected, the scientific practices employed and the research being performed. Yet, there are certain field activities whose context can be conveniently captured by producing and storing purely descriptive, short text annotations, i.e. micro-blog posts. In this article, we present the design and development of a cell phone application to enable field-based researchers to collect observational data and contextual information about field practices.

SDP 08.2 Approach

In developing tools for the capture of contextual information from field-based research, we attempt to address two specific challenges:

- **Interoperability.** A conceptual and technical discrepancy exists among available metadata standards used to collect and represent contextual data in sensor network research. Locally-specific metadata representation formats for datasets and field notes often fail to inter-operate. In research areas new to computer-based instrumentation, few generally accepted standards exist for description and annotation of resources such as sensor-collected ecological data.
- **Mobility.** Ecological field deployments using WSS necessarily involve in-field data creation, retrieval, and tracking. Understanding how highly variable field-based methods impact the resulting data products is essential to understanding and interpreting data on both short and long terms. Many field locations where WSS deployments take place do not have Internet or local area network connectivity. Thus, approaches to collecting information about data collection activities for WSS deployments cannot be solely web-based. Mobile applications and devices may enable field researchers to perform basic data collection and management operations when working in remote locations where Internet connectivity is lacking.

In addition to the aforementioned challenges, we note that the data (and related contextual data) collected and managed by scientists and engineers performing WSS-based field research are extremely heterogeneous in nature. This is due to the fact that sensor network research involves the interaction among researchers from a wide spectrum of domains (e.g., biologists, seismologists, electrical engineers, computer scientists) with differing scientific data

practices. Data handled by CENS field researchers differs not only in media types (text, images, audio, video, software) but also in the conceptual role that these data have within a broader data lifecycle.

SDP 08.3 System(s) Description and/or Experiments

This section describes our design and ongoing development of a mobile phone-based application for data collection and note taking in field research. CENS researchers who are engaged in field research normally work in small teams in specific locales using ad-hoc methods and heterogeneous data sources, and require flexibility to adjust research plans on the fly. Thus, in designing a cell phone application for this community, we have focused on enabling flexible and adaptive research. Our development is following the usability engineering lifecycle, which organizes the design process around two phases: requirements analysis and design, testing, and development.

The first phase, requirements analysis, involves setting up a user profile, performing contextual analysis of the users' main work tasks, setting usability goals, analyzing the platform capabilities and constraints (we are using the Windows Mobile 6.1 operating system for mobile phones as our development platform, and Samsung Omnia smart phones as our hardware for the initial stages of user testing), and creating general design principles. For our mobile application for field scientific research, our design principles were the following:

- 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 application as without: 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, thus our system must integrate with existing data and data tools.
- Produce data annotations that "live with" the data: Packaging data and its contextual information increases the usefulness, and consequently the value, of the data itself.

These principles stem from our analysis of field research activities in CENS. In recent sensor network deployments, scientists and engineers involved in field work collected many diverse kinds of contextual data such as information relative to the equipment used, participants of the project, annotations of various types, and a number of tasks performed during, before and after the field work. Some of this information is very domain-specific. Although disparate kinds of contextual data are being collected, there are certain categories of field activities that occur with regularity across all domains and practices of field-based research. This allows us to focus our design and development efforts on specific categories of field activities, yet leaving the system open for user appropriation.

The second design phase, design/testing/development, involves a number of iterative steps, including: conceptual modeling, conceptual mockups, setting screen design standards, prototyping, evaluation of the standards and prototyping, interface design, and evaluation of the interface design by gathering user feedback.

The focus of our development work thus far has centered around two common activities of field researchers: collecting tabular data and writing free text notes. To enable the collection of tabular data, we have built a web interface that allows a user to "author" data collection tables on their personal computer prior to going out in the field. The research can then upload the data tables as XML files to the mobile phone for use in the field. Once loaded on the phone, these data collection tables are customizable, in that they can be changed and re-used, which will facilitate data repeatability. Figure 1 shows a screen capture of a project home page, a data collection table, and on the mobile phone. The menus on the bottom shows the options available to the user, including adding an entry (data point), adding columns, adding/viewing notes, and seeing the geo-tagged items, which are discussed further below.

Moving to the note taking function, in all types of field work, researchers collect notes and annotations regarding their activities. The added value of the note taking functionality in our application is that the notes taken by the user are automatically date and time stamped and associated with a data table. The right-hand image in Figure 1 shows a screen capture of the note-taking interface.

As Figure 1 also illustrates, we have incorporated a "geo-tagging" feature into the application. Users can geo-tag data points and notes by checking the "Geo-tag this" box (shown in the right-hand screen). The geo-tag function pulls the latitude and longitude of the user's current location from the phone's built in GPS receiver and links the coordinates to

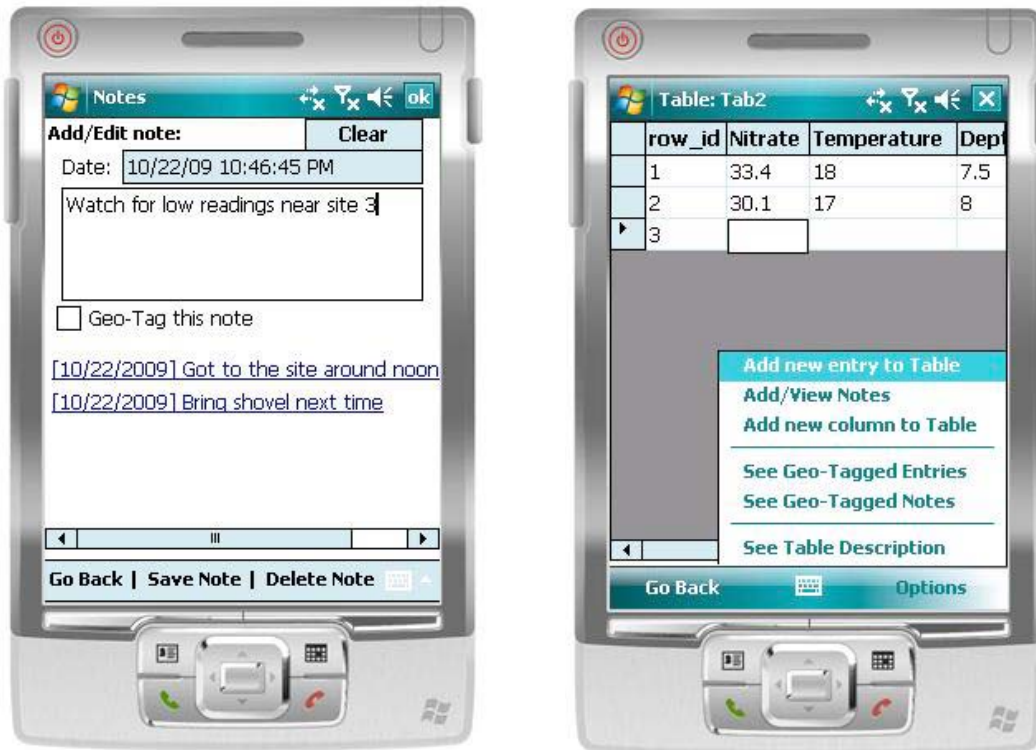


Figure 1: Screenshots from left to right: data collection table and note-taking interface

the data point or note that was tagged. The user can then filter the display to view the data points and notes that have been geo-tagged, though in many situations the GPS coordinates are of little meaning or use in and of themselves, so we are exploring ways to visually present the geo-tagged information, for example through a map interface. Also, even devices that have GPS capability may not be usable in field locations because of physical obstructions. In these cases, our note taking application provides a quick and efficient tool to associate a particular self-described location with the current timestamp. Although geographically less precise than GPS, these descriptions might provide some context when GPS data are entirely missing. It is straightforward to envision how these functionalities could be extended to incorporate micro-blogging tools and services. Updates could be sent out regarding the progress of data collection. The notes could be sent to a Twitter feed, with date, time, and GPS stamps.

SDP 08.4 Accomplishments

In this section, we present some results from preliminary user tests of our application. Our first round of pilot users consisted of three scientists with extensive experience in performing environmental and ecological research in field settings. The pilot tests took place indoors in a lab setting, and focused on the usability and utility of the data collection and note taking applications described in the previous section. Additionally, we asked the users about their experience in using the application, as well as how the collection of data and contextual notes might have been enhanced by the use of a dedicated micro-blogging feature.

The pilot test protocol consisted of five tasks aimed at testing the main functionalities of the system, and how easy the application was to learn and use. First, we had the users test the web interface that allowed them to create data collection tables, and second, we had them test the application using these tables. This involved opening the tables, taking data, customizing the tables by creating new columns, taking notes, and finally exporting the data. We did not help users to use the application because we wanted to mimic the real-world environment as much as possible. We asked the users to think aloud as they completed the tasks in order to gain a better understanding of their thought processes. The feedback from our pilot users highlighted both the positive and negative aspects of our prototype system. Users liked the ability to create their own data tables, and the ability to customize the tables during the data collection process, by adding new columns. They also liked the mobile aspect of the application, in that the small size of a mobile phone will not be an encumbrance as they perform their research activities in varied field settings.

The small size of the application had both positive and negative aspects. The small screen size of the phones might prove to be hard to see in some situations, specifically mentioning button size and text size as potential issues. We

also identified individual pages and functions that were either hard to use (such as the data import function) or did not work as we had anticipated, allowing us to see where the application needed to be streamlined for a smoother user experience.

The pilot users also responded positively to our questions about micro-blogging. One user said that having a microblogging system that collected notes as tweets or text messages and displayed them in a database or on a website would be very useful. The pilot users also indicated that a micro-blogging functionality in our application would be most useful in situations that involved collecting information about events. More specifically, sending notes out as micro-blogs would be useful in providing notifications to other members of the research team or to the general public that a certain event has taken place, or that important thresholds have been achieved. The users also mentioned that attaching GPS or other location information to the micro-blogs would increase their information utility.

The pilot test described here is clearly limited in scope and scale, but it has been useful in getting preliminary feedback from field scientists on key functionalities of our application, and our ideas for next steps. The users helped us to identify where our next development thrusts should go, and show that implementing a micro-blogging service as part of our application would be useful both as an individual and a team research tool.

SDP 08.5 Future Directions

A fuller evaluation of the application will be conducted at a later date as the project matures. The fuller test will be conducted in real-world field settings, and will test how well the application meets the five goals outlined above.

Our ongoing work suggests that micro-blogging has the potential to be widely adopted as an annotation tool by field researchers for at least two reasons. First, micro-blogging is already a widely popular and used form of communication on the web, especially on social networking sites. By providing researchers with tools that they are already familiar with, we envision to obtain wide adoption by the intended users. Second, the broadcasting of micro-blogging updates can be customized to be private (e.g., for personal use), group-restricted (e.g., for use within a given project) or public (available to anyone). By letting researchers adopt different levels of micro-blogging privacy, we envision that they will engage in patterns of interactivity (e.g. sharing annotations, subscribing to each other annotation feeds, etc.) that are ultimately beneficial for the pursuit of collaborative field research. In the long run, widespread use of micro-blogging and related annotation tools will enrich data digital libraries with important contextual information, thus enabling reuse and interpretation of data collected in field activities.

SDP 09 Object Reuse and Exchange; RDF data modeling.

SDP 09.1 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) and the Resource Description Framework (RDF) data models to describe, publish and share aggregations of information objects produced at different stages of the scientific lifecycle in environmental sensing research. A conceptual implementation of such system for two specific CENS case scenarios has been discussed in a recent scholarly publication. We are now completing this work by developing a lightweight testbed to publish and share such sensor-specific aggregations to the Semantic Web.

SDP 09.2 Approach

The work presented here builds on previous research in which we developed a conceptual model of the CENS scientific lifecycle. 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 09.3 System Description

In a recently published article (Pepe et al., 2010), we have extended the notion of the scientific lifecycle for ecological sensing data. We demonstrate a conceptual implementation of OAI-ORE to represent the scientific life cycles of two embedded networked sensor applications: one in the field of seismology (the Southern Peru, formerly Middle American Subduction Experiment (MASE)) and one in the field of environmental science and water contamination (the Networked Info-Mechanical Systems (NIMS) deployment in the San Joaquin River).

In particular, we have proposed a technical system implementation of the OAI-ORE protocol to enable description of the processes and practices that take place in a typical CENS scientific lifecycle. Via this implementation, we plan to

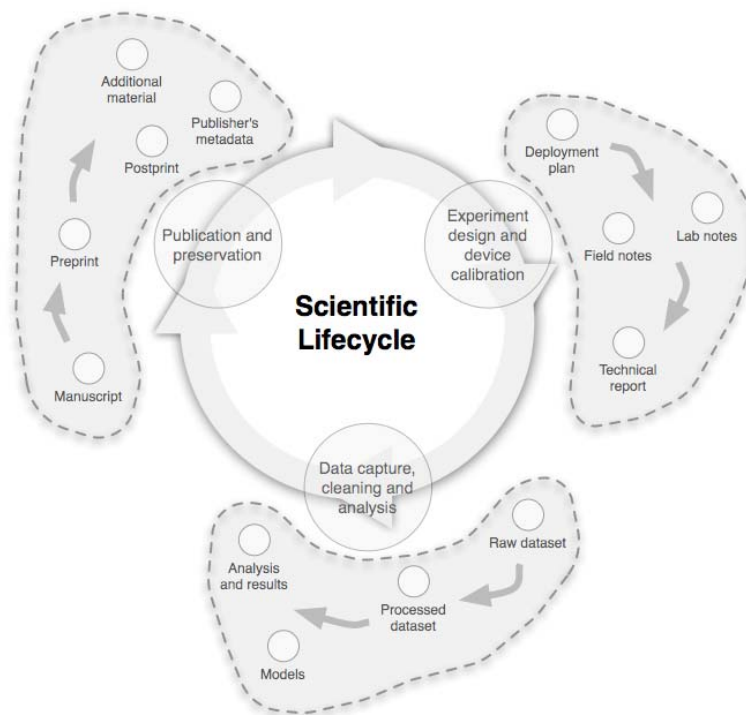


Figure 1

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 and at other locations on the web. Figure 1 depicts a conceptual enhanced version of the scientific lifecycle for a typical environmental sensing application.

One of the case studies presented in the article – a seismic sensing deployment in Southern Peru in 2008 – is based on ethnographic research of CENS field deployments and interviews. We have reconstructed the scientific lifecycle of this deployment and generated conceptual aggregations of its information products in OAI-ORE (Figure 2).

By establishing relationships between publications, data, and contextual research information, we illustrate how to obtain a richer and more realistic view of CENS scientific practices. That view can facilitate new forms of scientific research and learning.

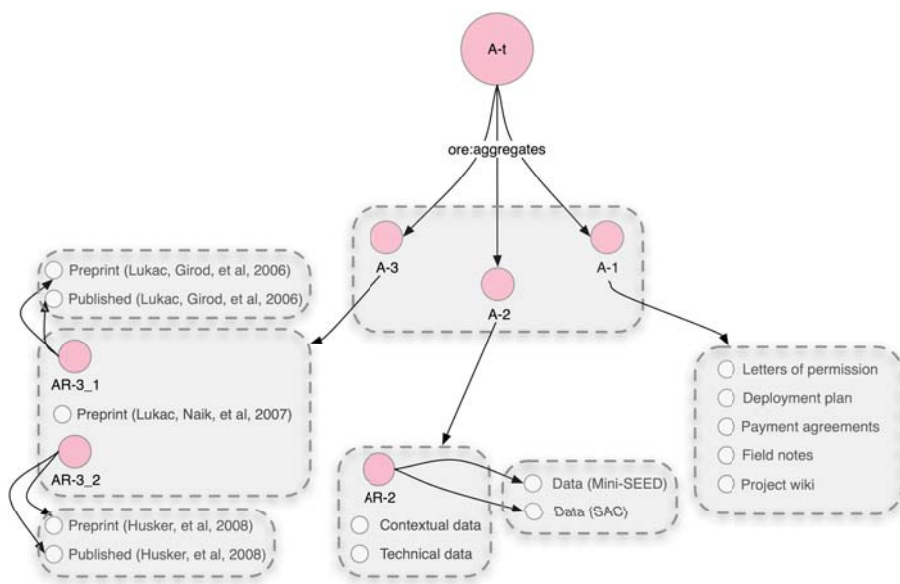


Figure 2

SDP 09.4 Accomplishments

Specific accomplishments during the reporting period:

Developed and disseminated model of the CENS data lifecycle.

- Demonstrated the potential use of RDF to model information object relationships for network analytic applications (Pepe and Rodriguez, 2010)
- Presented this and related work at national and international conferences
- Comprehensive journal article published: (Pepe, Mayernik, Borgman & Van de Sompel, 2010) Pepe, A., Mayernik, M., Borgman, C. L. & Van de Sompel, H. (2010). From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web. *Journal of the American Society for Information Science and Technology*, 61(3): 567–582.

SDP 09.5 Future Directions

By the end of this year, we plan to develop and make available a demonstration prototype for the manual generation and publication of ORE aggregations on the Semantic Web using RDF for sensor resources. This will allow partner scholarly and scientific repositories to harvest aggregated information about CENS research in a structured format.

SPD 10 DataNet: Data Conservancy (UCLA-DC).

SPD 10.1 Overview

The Data Conservancy is a five-year project lead by Johns Hopkins University, funded by NSF's DataNet initiative by the Office of Cyberinfrastructure. A primary goal of DC is development of a repository that will archive data and facilitate collaborative access for a range of sciences. Its design will be based upon social science research of data curation and collaboration practices for science communities who are likely users of the DC repository. UCLA's researchers, administered through CENS, will conduct research on practices and data curation requirements for astronomy and astrophysics. The outcomes of their research will contribute to the design of the Data Conservancy system architecture.

SDP 10.2 Approach

UCLA Data Conservancy (UCLA-DC) is coordinating with social science teams at The Center for Informatics Research in Science and Scholarship (CIRSS) at University of Illinois, Urbana-Champaign, and the National Center for Atmospheric Research (NCAR), who are studying data practices and needs for a range of science communities expected to contribute their data to the Data Conservancy and who will be users of DC's products and services. UCLA-DC will extend their current work on scientific data practices and data curation requirements in embedded sensor networks to the fields of astronomy and astrophysics. We will examine initially the Sloan Digital Sky Survey (SDSS), and its relation to two subsequent sky survey projects: the Large Synoptic Survey Telescope (LSST) and the Pan-STARRS project. Also examined are the Infrared Processing and Analysis Center (IPaC), the International Virtual Observatory Alliance (IVOA) and Space Telescope Science Institute Archives (STScI) for relevant issues in data practices and metadata standards for astronomical objects and digital archives. Core questions for the three astronomy projects (SDSS, PAN-STARRS, and LSST) are focused on their design, structure, usage, and data practices, in support of data curation in particular. UCLA-DC will examine their history of development, core practices of data management and curation, hurdles overcome and remaining, and lessons learned that are instructive for related projects within and outside astronomy. Investigators will determine which forms of data are used, which are selected for sharing and curation and which are discarded, how they are curated, and expected future uses of curated data. We will employ a range of qualitative methods to address these goals including analysis of documented history of projects, oral histories and interviews with scientists, developers, and data managers, ethnography including observations of sites, and social network analysis of the involvement of key participants over time.

SDP 10.3 System(s) Description and/or Experiments

One central activity of the project will be the building and analysis of a relational database of documentation on project sites. The database will track project history, funding, personnel, timelines, and relationships among projects. The database will also incorporate for analysis data generated by our other methods. The database is built on an SQL network platform with a user interface operating in FileMaker Server.

SDP 10.4 Accomplishments

Specific accomplishments during the reporting period:

Borgman launched new graduate course on "Data, Data Practices, and Data Curation."

- Several meetings with main project partners, including project PI, SDSS director, and project partners at CIRSS and NCAR.
- Development and initial deployment of a relational database and document management system for organizational, ethnographic, and social network analysis.
- Established a relationship with the Caltech library, which helps manage large astronomy/astrophysics data centers and digital libraries, and who will aid in our contacts at Caltech and JPL.
- Had a facilities tour and conducted several interviews with personnel and astronomers at Caltech's Infrared Processing and Analysis Center (IPAC), to develop comparative materials for digital data management.
- Meetings with Microsoft Research, including VP Anthony Hey, to explore partnership directions with UCLA-DC and the DC project.

SDP 10.5 Future Directions

Our full interview and site visit schedule will begin in Summer 2010. Interviews at LA-area astronomy/astrophysics institutions will continue. Concurrently, we will be developing and populating our database with information about SDSS and other project sites, and extending its integration with our interview and ethnographic data, and social network analysis capacities. We will be coordinating activities with CIRSS, NCAR, and Microsoft partners.

All empirical results, use cases, and accounts of scientists' data practices and conceptualizations of data will be provided to the Data Conservancy Data Practices and Data Concepts groups with regular working sessions for joint interpretation. The UCLA-DC team will also collaborate with DC researchers at CIRSS and NCAR to integrate related findings, sharing results and interpretations of findings through the project wiki, conference calls, and scheduled face-to-face meetings. The overall analysis will produce a taxonomy of data practices and data attributes for assessment of curation needs and to facilitate curation activities.