

2.7 Research Testbeds and Data Practices (DATA)

These projects are both testbeds and empirical research; the common theme is collaboration, both local and distributed. We are exploring multiple aspects of cyberinfrastructure requirements for embedded sensor networks and other types of multi-disciplinary science and technology research. Funding sources include NSF, Microsoft, CENS, and the University of California.

Monitoring, Modeling, and Memory is a distributed collaboration with partners at the University of Michigan, University of Pittsburgh, and Georgetown University. We are comparing data practices across multiple distributed sci-tech collaborations, including CENS, Long Term Ecological Research Network (LTER), and the Earth System Modeling Framework. This is largely a social science collaboration of investigators studying science. We are in year three of three.

The Data Conservancy, which is a very large distributed collaboration based at Johns Hopkins University, addresses the curation of scientific observations. It is a partnership of scientists, social scientists, and technologists to build systems and to study data curation in multiple fields. Our focus is on astronomy, which is outside the realm of CENS per se, but draws upon the theory and methods applied to other studies of CENS data practices. We are in year two of five.

Object Reuse and Exchange is a technology project that builds upon our earlier work in CENS data practices, extending it to deploy new technical standards to represent and to link research objects. We partnered with leads in ORE development at the Los Alamos National Laboratory. This project is in year 4 of 4. Funding for this research came primarily from Microsoft Research.

The CENS Deployment Center captures knowledge of CENS research activities and makes those records reusable for future deployments by the same and other CENS teams. The eScholarship project captures CENS publications. The Data Discovery Library and the CENS Annual Report System both were outcomes of these projects, which began with a seed grant from CENS funding. While CENS DC per se is ending, the follow-on projects continue, and a grant proposal is in progress to transfer stewardship of the CENS data and publications to the UCLA library.

The mobile computing project, also funded by Microsoft, developed a scientific data collection module for the Microsoft phone platform. While technology development ended when Microsoft made a major platform shift that was not forward compatible, we gathered useful empirical data on microblogging for scientific applications.

The multiple CENS testbeds are exemplars of the collaborative cyberinfrastructure built by CENS teams. These testbeds yield empirical data and are available to other teams to extend their work.

Theses and Dissertations

MMM, Deployment Center, and eScholarship: These three projects together are the basis for the dissertations of Matthew Mayernik, who will complete his degree in June, 2011, and Jillian Wallis, who defended her proposal in March, 2011, and plans to finish in June, 2012.

Object Reuse and Exchange: Alberto Pepe, who completed his dissertation in June, 2010, extended the ORE project with additional study of social networks in CENS.

Mobile scientific data collection: Research done jointly by Mayernik and Pepe contributed to both of their dissertations.

Data Conservancy: David Fearon's master's degree project is based on his work with astronomy data from this project.

Awards and Prizes

In this reporting year, our team has received multiple awards:

- Matthew Mayernik: Best Dissertation Proposal Award, UCLA Department of Information Studies
- Alberto Pepe: Best Dissertation Award, American Society for Information Science and Technology
- Jillian Wallis: Team lead, Winner of Student Design Competition, American Society for Information Science and Technology
- Laura Wynholds: Inaugural Best Student Paper Award, International Digital Curation Conference

DATA 01 Monitoring, Modeling, & Memory: Dynamics of Data and Knowledge in Scientific Cyberinfrastructure

Team Members

- Christine Borgman, Faculty, PI*
- Geof Bowker, Faculty, PI
- Paul Edwards, Faculty, PI
- Steven Jackson, Faculty
- Matthew Mayernik, Graduate Student
- David Ribes, Faculty
- Elizabeth Rolando, Graduate Student
- Leigh Star, Faculty
- Jillian Wallis, Graduate Student*

** Primary Contact*

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.

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:

- 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?for the CI vision therefore include:

Experiments

To answer these questions, we are in year three 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.goals.

Accomplishments

At the end of summer we wrapped up our second year with a research retreat at the University of Michigan, 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 second retreat of three to support an iterative cycle of comparison of the cyberinfrastructure projects. Our previous round of data collection has completed, and we are simultaneously performing analysis on this round of interview data and planning focused data collection efforts to dive deeper into the specific areas of metadata practices and temporal rhythms that emerged as topics of interests from our more general interviews. A Masters student from the Department of Information Studies was hired to assist in data analysis and collection.

A major difficulty hampering cyberinfrastructure research is a lack of theoretical frameworks, making it difficult for research to progress past case studies. We are engaging with other CI researchers to develop larger research trajectories that can help us move forward. To this end we engage in discussion of this topic within our group regularly and are bringing the discussion to a larger audience at conferences, such as the American Society for Information Science & Technology 2010 Annual Meeting where we held a very successful panel to discuss just this topic. In order to continue this research for an additional year we have applied for a no-cost extension from the NSF.

Future Directions

During the upcoming year we will hold another research and writing retreat to bring together collaborators and coordinate our efforts. We will collect focused data based on the protocols developed during this year, analyze this data and publish the results. We look forward to attending the Joint Conference on Digital Libraries 2011 Meeting, the 2011 Annual Meeting of the Society for the Social Studies of Science, and the American Society for Information Science & Technology 2011 Annual Meeting, where we will continue to engage the community on taking CI studies beyond the case study.

DATA 02 DataNet: Data Conservancy (UCLA-DC).

Team Members

- Christine Borgman, Faculty, PI*
- David Fearon, Graduate Student*
- Sharon Traweek, Faculty
- Laura Wynholds, Graduate Student

* Primary Contact

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.

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 is studying scientific data practices and data curation requirements for the fields of astronomy and astrophysics. We are examining 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. Additionally, we are conducting a series of interviews with astronomers and astrophysicists, including university faculty, postdoctoral and graduate students, and data center staff. Core questions for astronomers are focused on data sources, usage, and preservation. 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. In addition to interviews, we are conducting a range of qualitative methods to address these goals including analysis of documented history of projects, observations of labs and workstations, and social network analysis of the involvement of key participants over time.

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.

Accomplishments

Specific accomplishments during the reporting period:

- Hired GSR David Fearon for a full-time post-doctoral staff position for the project year 2 with Jim Gray Gift funding from Microsoft Research
- Acquired and began analysis of the Sloan Digital Sky Survey listserv archive held by the Data Conservancy project.
- UCLA GSR Laura Wynholds met with the Data Conservancy CIRSS project partners at UIUC, before iConference at Urbana-Champaign. Contributed to collaborative partnership between UIUC and UCLA.
- Completed 25 interviews, most with faculty astronomers in California, including senior astronomers on prominent projects. These complete the "round-one" interviews, 32 total.
- Completed initial round-one coding and analysis of interviews.
- Round-one Analysis products included two conference posters, a white paper for meetings with Data Conservancy project partners, and a conference paper currently in review.

- Co-PI Sharon Traweek spent much of August at the KEK High Energy Accelerator Research Organization in Tsukuba, Japan, engaging with several astrophysicists and gathering background on data curation in Japan.
- Held two half-day meetings with astronomer Alyssa Goodman, discussing astronomy data curation, publication data linking.

Future Directions

We are beginning our second round of interviews in spring 2011, continuing through summer, primarily with Southern California astronomy/astrophysics institutions, will continue. Interviews will focus in particular on the preservation of recently published data, linking of data to publications, and integration of archived sources for data-intensive research. Concurrently, we will be continuing our document and listserv archive about SDSS development and other project sites, and extending integration of our project databases with our interview and ethnographic data, and social network analysis capacities. We will be coordinating activities with CIRSS, NCAR, MBL, 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, NCAR, and MBL 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.

DATA 03 Object Reuse and Exchange; Studies of scientific collaboration.

Team Members

- Christine Borgman, Faculty*
- Alberto Pepe, Graduate Student, PI*

* Primary Contact

Overview

The work briefly presented here reports on two parallel and inter-related research endeavors, which are now concluded. The first endeavor involves the design and development of tools to allow efficient reuse and exchange of information objects resulting from embedded sensor network research applications. The second endeavor stems directly from the dissertation research of the project's PI and deals with the study of scientific collaboration networks at the Center for Embedded Networked Sensing.

Approach

The Object Reuse and Exchange portion of this project builds on previous research in which we developed a conceptual model of the CENS scientific lifecycle (Figure 1). 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.

The study of CENS scientific collaboration was developed from 2006 to 2010 and was reported in an unpublished dissertation and forthcoming publications. By use of survey research and network analysis, this part of the project examined the collaborative ecology of CENS in terms of three networks of interaction: co-authorship of scholarly papers, communication activity on mailing lists, and interpersonal acquaintanceship.

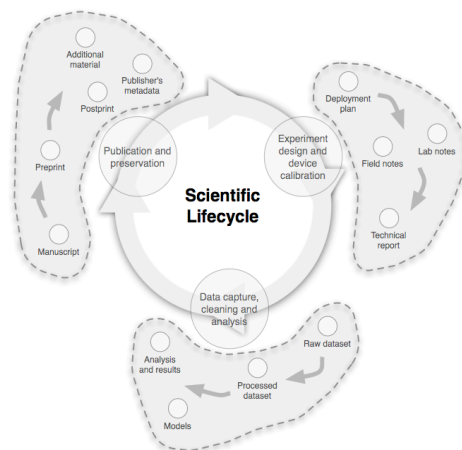


Figure 1. A conceptual model of the CENS scientific lifecycle.

System Description

As part of the Object Reuse and Exchange project, we identified three major digital resources across the CENS data life cycle: a) information about deployments, b) sensor data and c) scientific publications. In published work, we adapted the OAI-ORE data model to describe, publish and share aggregations of information objects produced at different stages of the CENS scientific lifecycle and across the three aforementioned digital repositories. In submitted work, we proposed the use of microblogging to document field-based research and the linkages that exist between deployment information and related scientific artifacts.

The study of scientific collaboration exposes the topology, structure, and evolution of CENS networks in relation with the disciplinary and institutional arrangements of the center. Findings of this research point to the importance of interpersonal relationships for accomplishing scientific work in distributed environments. Network analyses reveal that structural communities in the co-authorship and acquaintanceship networks overlap considerably. The community structure of the acquaintanceship network is shown in Figure 2.

Accomplishments

The accomplishments for the reporting period are limited to the editing, preparation, and submission of the scholarly papers related to this research project (currently in submission).

Future Directions

This project is discontinued as of June 30, 2010.

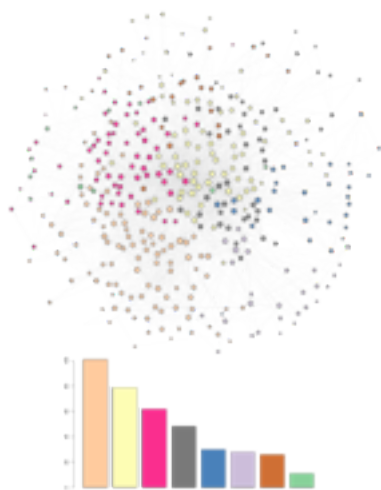


Figure 2. Pictorial representation of the acquaintanceship network at CENS. Structural communities in the coauthorship network detected according to the spinglass algorithm. Node color represents structural community membership. Node diameter represents betweenness centrality. Associated histogram describes the frequency distribution of each community.

DATA 04 CENS Deployment Center and Metadata Repository

Team Members

- Christine Borgman, Faculty, PI*
- Matthew Mayernik, Graduate Student, PI*
- Jillian Wallis, Graduate Student

** Primary Contact*

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 and CENS metadata repository projects are how to facilitate knowledge transfer in the collaborative CENS research setting, and CENS data can be made available for secondary uses outside of CENS.

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 has been implemented as part of the CENS annual report generating system.

Approach

Our goal for the CENS metadata repository is to enable CENS data to be discovered by potential users. Among the data resources collected by CENS researchers are images, audio files, physical samples, and numeric data in both digital and analog form. These data resources are distributed around community, lab, and individual computer systems. Some CENS research data and documentation are available online through lab websites, but large portions of these resources reside in protected computer systems, personal laptops, file cabinets, and even in refrigerators. Collecting and integrating all of the Center's data into a single system would be prohibitively expensive and time consuming, even if investigators were willing to release them. Thus, we are collecting descriptions of data sets created and used by CENS researchers instead of collecting the data themselves. We consider these descriptions of data sets to be metadata about CENS data. The CENS metadata repository is designed to enable potential data users to discover what CENS data exist, to determine whether those data may be useful, and to learn how to acquire data of interest.

System(s) Description and/or Experiments

The CENS metadata repository uses a set of metadata elements drawn from the Dublin Core Metadata Schema. We then re-named the Dublin Core elements in order to reflect terminology that is familiar to the scientists, engineers, and computer scientists in CENS. The metadata repository asks CENS researchers to describe their data sets using the following metadata fields: title, collection start and end date, location, people who contributed to data collection, data type (image, numerical, etc.), research question, variables, data collection process, data format, permissions, funding source(s), keywords, location of data set, related publications, and related CENSDC record.

In the past year, we have rolled out the initial version of the metadata repository as part of the 2010 online CENS annual reporting system. Following the completion of the 2010 CENS annual report to the NSF, we conducted user tests of a prototype of the second version of the metadata repository. In these user tests, we asked CENS researchers to describe their data using the metadata fields listed above. After the testers completed the test, we asked targeted questions about their experience in performing the task. Post-test questions included asking the researchers which metadata fields they felt were the most and least useful in describing their data, what additional fields might be necessary, and what benefits (if any) they feel that they receive from creating this metadata, among other questions.

Accomplishments

The initial version of the CENS metadata repository was a mixed success. Of the roughly 65 project reports submitted as part of the 2010 CENS annual report, which is estimated to be a 90-95% response rate from all of the CENS research groups, 11 datasets were submitted. Four of these 11 datasets were submitted by the data practices group. The other seven datasets consist of four from the Participatory Sensing, and one each from Contaminant Transport

The screenshot shows a web form for dataset metadata submission. It is titled "Dataset Information" and contains the following elements:

- Title:** A text input field with an asterisk indicating it is required.
- Dates of collection:** Two text input fields for "Start Date:" and "End Date:". Below them is a note: "Please format dates this way: MM-DD-YYYY."
- Data collection site:** A text input field.
- Contributors:** A section with the instruction "People who contributed to data collection, choose primary contact person:" and a button labeled "Add/Change Contributor(s)".
- Data Type:** A section with the instruction "Please select all that apply." followed by a list of checkboxes and their corresponding descriptions:
 - Events (capturing data in response to specific events, such as seismic events, algal blooms, etc.)
 - Geo-spatial (has GPS or other lat./long. information)
 - Image (digital or film photos)
 - Interactive Resource (web forms, applets, multimedia objects)
 - Moving Image (movies/videos)
 - Numerical (tables or files of numbers)
 - Physical Object (physical samples or specimens)
 - Software (source files, executables, scripts)
 - Sound (audio recordings)
 - Text (words or textual narratives)
 - Time Series
- Research question/why the data was collected:** A text input field.

Figure 1. Screenshot of the metadata submission form that is included in the 2011 CENS annual reporting system. Only six metadata elements are shown.

and Management, Multiscale Actuated Sensing, and our statistics partners in the Statistics and Data Practices group. No datasets were reported by other science groups such as Seismic, Aquatic, or Terrestrial sensing, or by the other computer science and engineering groups in this first round.

Based on our previous research, it is likely that at least one dataset fitting the reporting criteria could be associated with each of 70 or so individual research reports. This low response rate can be attributed to a variety of factors. First, the CENS solicitation was ambiguous about whether reporting datasets was mandatory. The solicitation language was carefully written to reflect the facts that NSF has specifically requested that CENS report datasets, but that datasets are not a mandatory section of the NSF official reports. Second, the reporting process changed in several ways this year, creating some confusion. This is the first year in which CENS used an online reporting platform, where each report

section is entered into a web form. In prior years annual reports were submitted manually, i.e., documents in a template sent by email. The template included only the sections mandated by the NSF (description of research, people, publications, etc). Third, the new Annual Report System impeded dataset submission in two crucial ways: 1) the new system did not require the majority of fields, resulting in gaps in the dataset, publication, and general report metadata, and 2) the user interface was difficult to use. We anticipate that these problems can be reduced significantly in the second version of the Annual Report System. The system was implemented somewhat prematurely, lacking some planned interface features, due to the hard deadlines for NSF annual reports.

A fourth reason for the low response rate, and one of particular interest for our data practices research, is the diffuse responsibility for datasets. Multiple research groups may be involved in the production of any given dataset. Conversely, any given deployment may result in multiple datasets used by different researchers and teams. Other data are produced in laboratories or in simulation studies. Once collected, different individuals may analyze, manage, and report these data in publications. It is often unclear which individual has ultimate responsibility for any given dataset or who should take responsibility for including it in an annual report. The sections of annual reports often are delegated to graduate students who are the leads on specific projects. Faculty team leaders review the reports and synthesize results across research areas of CENS. The determination of who has ultimate responsibility for a given dataset is among our research questions.

Despite the low response rate for datasets in the first iteration of the CENS metadata repository, the reported datasets were listed in the 2010 CENS annual report to the NSF. This is the first year that CENS has reported datasets in an annual report, which is a significant step forward in increasing the availability of CENS datasets for secondary uses. We hope that the user tests we conducted following the completion of the 2010 annual reporting cycle will help to increase the response rate during the 2011 reporting cycle. The user tests helped us to identify where confusion might occur as researchers fill out the metadata form. The user tests also allowed us to conduct outreach to CENS research groups that did not submit metadata for data sets in the 2010 annual report. Figure 1 shows a screenshot of the metadata submission page that is being used in the 2011 annual report submission system. For the 2011 annual report, we have corrected misleading or unclear terminology, and provided more flexibility in the ways that researchers can respond to certain metadata fields.

Future Directions

In the upcoming year, we will continue to assess the success of the metadata repository. We will evaluate the responses that we receive as part of the 2011 CENS annual report, looking at the metadata submission response rates and the metadata submissions themselves. This evaluation will give us a good indication of how our user testing and outreach efforts can impact submission response rates for community data or metadata repositories, both

within and outside of CENS. Depending on this evaluation, we will conduct more user tests and interviews to determine what might be improved for the 2012 annual reporting cycle.

Making CENS data sets visible and discoverable on the web is a main goal of our research. Thus, the second major thrust of work for the upcoming year will be to develop a web presence for the metadata submissions on the CENS website. This will involve developing a data structure that is compatible with the Open Archives Initiative Protocol for Metadata Harvesting, and developing web pages that display the metadata submissions in a flexible way. We will also conduct evaluations of this effort by recording download/viewing rates and by getting feedback from CENS researchers on our efforts.

DATA 05 Mobile Scientific Data Collection

Team Members

- Christine Borgman, Faculty, PI*
- Matthew Mayernik, Graduate Student*
- Alberto Pepe, Graduate Student

* Primary Contact

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.

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). Our proposed approach 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.

Approach

What tools can we provide scientists to support collection of contextual data and documentation of the variegated artifacts that are produced in the scientific information lifecycle? We propose the use of microblogging, or short text annotations, as a mechanism to document the context and data collection practices of field-based research applications. Our approach to this topic was to investigate researchers’ current use of microblogging, and to design and develop a cell phone application to enable field-based researchers to collect observational data and contextual information about field practices. In past reports we have described our design considerations, prototype development efforts of a handheld application for collecting contextual data, and pilot test results. In this report, we describe the results of interviews with researchers involved in in-situ ecological monitoring and related sensing research about how microblogging can facilitate the capture of contextual information for in-situ monitoring and related WSS-based field research.

System(s) Description and/or Experiments

We performed an exploratory survey among field researchers to determine the current extent of adoption of Twitter and similar microblogging platforms, whether it is for research or social purposes. We conducted interviews with eight scholars of different ranks (research staff and graduate students) involved in field research of various kinds (ecology, biology, environmental science, and urban sensing). Interviews were conducted in person or via email in a semi-structured format. The questions asked throughout the interview are the following:

- Have you ever used a micro-blogging service (e.g. Twitter)? If yes, what have you used micro-blogging services for? Do you use micro-blogging for research purposes?
- Do you know of other researchers in your discipline who use micro-blogging for research purposes? If so, what do they use micro-blogging for?
- Do you think that implementing a micro-blogging service on top of existing data collection tools/devices would assist you in your field work in any way? What specific information would micro-blogging facilitate to collect in the field?

- Do you think that micro-blogging could enable exchange of information (i.e., research-based conversations) with fellow researchers during field activities? If so, in what ways would that be useful for you?

Accomplishments

When asked about their micro-blogging practices, nearly all interviewed researchers said that they either do not use Twitter (or any similar micro-blogging platform), or if they do, their use is limited solely to social interactions. Only one researcher actively employed Twitter both for social and research purposes. This use is specifically limited within the scope of Project Budburst, a citizen science mobile sensing project aimed at collecting data on phenology and climate change. Project participants contribute to this project by performing multiple observations of the same plants, documenting “phenophases” such as their first leaf, the first flower, and the first fruit of trees, shrubs, flowers, and grasses. There are multiple ways to make and record budburst observations, including sms text, a smartphone application, and Twitter. On Twitter, users mark tweets about observations with a given hashtag (@budburst). The application scrapes the tweet corpus daily and a stream of observations becomes available to users and researchers on the project. The same researcher also indicated another research use of Twitter, which is not directly field-based, but that is related to the aforementioned project. In what they jokingly referred to as “opportunistic” participatory sensing, they continuously search the public Twitterverse for phenology-related keywords, such as “spring”, “bloom”, “blossoming”, etc. They log tweets containing these key terms and their geotag, if available. Analysis of this corpus allows them to correlate user-contributed Twitter events with life cycle events in plants.

All the researchers that do not use Twitter also indicated that they do not know of any colleagues who do use it for research purposes. Some researchers were aware, however, of some field-based research applications that use aspects of blogging or micro-blogging as their data ingest mechanism. One researcher mentioned the existence of a Facebook application for beach contamination source tracking. In the realm of urban sensing, a researcher mentioned a project of a colleague who uses the distribution of tweets across cities to look at urban landscapes and urban density. Nearly all interviewed researchers indicated that the current most common and beneficial use of Twitter for research purposes has to do with scientific communication, dissemination of scientific results, and gaining visibility in one's scientific community.

Researchers indicated a number of ways by which microblogging services could assist their field work practices. One researcher suggested the use of Twitter feeds for automatic notification of instrument failures. Device failure is a problem that spans across all mobile sensing applications. The use of a Twitter feed, in addition to email notifications, would allow researchers to instantaneously subscribe to multiple devices, re-broadcast (retweet) given failures to a broader/different research community, and publish notes and annotations concomitant with the failure messages, for public viewing. One respondent mentioned the potential of microblogging platforms to foster public awareness and interest in scientific research. For example, it was suggested that Twitter feeds could be useful to publish water quality data about particular beaches. This in turn, has the potential to serve the research community---other researchers might be interested in harvesting, storing and reusing those data---as well as the public---citizen scientists and beachgoers may find interest in the data and in the processes by which those data were collected.

Some researchers pointed to an important shortcoming of Twitter: its inability to function in field locations in which Internet connectivity is not available. As discussed above, the mobile and unpredictable nature of most scientific field research implies that Internet connectivity may be lacking for extended periods of time. If Twitter had to be used as a note-taking tool for the collection of contextual data, how would it behave in the absence of Internet connection? The doubts expressed by interviewed researchers point to the importance to employ microblogging services with offline capabilities when performing field experiments. With regard to the last question of the interview, many researchers were unsure whether microblogging is a viable platform for exchange of research-based information during field activities, for two main reasons: the potential lack of connectivity in field settings and the public nature of Twitter. The connectivity issue was raised as a major cause for the inability to communicate while in the field. Other researchers were worried about the public nature of Twitter, and thus the issue of publishing raw, non-certified field data into the public. Although Twitter does allow accounts to be private, i.e., to be accessible only by a selected group of contacts, researchers expressed the need to make openly available certain sections of their Twitter feeds, while restricting others.

In sum, all researchers maintained to be somewhat familiar with Twitter, although only very few employ microblogging services actively. Some of them use Twitter for promotion of their scientific work and to network with other scientists. Although most interviewed researchers consider feasible and convenient the use of Twitter to support field-based research, only one of them has personally experimented in this regard, collecting field data about phenophases and climate change. Two major shortcomings of Twitter were put forward: its inability to work in offline mode and its limited range of options on access privileges.

Future Directions

While our interview and survey results illustrate how further development on our prototype application can extend and refine the ways that microblogging tools can help collect contextual data in in-situ sensing projects, we have stopped current development efforts. Our initial prototype was developed for cell phones that use the Windows Mobile 6.1 operating system. However, Microsoft has announced that its newly developed Windows Mobile 7 operating system will not support applications based in Windows Mobile 6.1. Thus, we would have to completely re-write our application in order to and enable it to be used on future Microsoft-based cell phones, which would be necessary to deploy and test our system more widely. Because of limited resources, we do not have the capability currently to conduct such a re-development. Thus, we are focusing on writing up our prototype and pilot work, and identifying where our lessons learned can be applied to similar projects in the future.

DATA 06 eScholarship Repository

Team Members

- Christine Borgman, Faculty, PI*
- Jeff Goldman, Staff
- Jillian Wallis, Graduate Student*

* Primary Contact

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. As CENS winds down we are looking towards the long-term sustainability of the eScholarship Repository and how we can take what we have learned with developing a repository for CENS and apply it to the larger institutional community.

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.

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 generating social network analysis data. Work within this project includes the regular upload and maintenance of materials in the eScholarship Repository, as well as finding new ways to make submitting publications easier or better incentivized.

Accomplishments

The work being done within this project received attention earlier this year from a UCLA campus-wide initiative addressing the concerns of faculty needing to draft data management plans for NSF funding proposals. As one of the few projects at UCLA where work was being done to tie together data collected during research and publications, we are uniquely poised to present a solution that could be scaled up to meet the needs of researchers across UCLA and the broader UC System. We submitted a proposal detailing how we would like to further develop the approach we are taking at CENS and how it could be scaled up to meet the needs of the community beyond CENS. CENS is an excellent testbed for this work given the diversity of CENS members which span from computer science and engineering to marine biology and seismology to statistics and education. To develop a publication and data repository that works for all these members we have need to leverage the common practices, which tend to fall within reporting to institutions and funders. Systems that assist researchers in their reporting tasks, such as submitting annual reports to funders or materials to review/tenure committees can then take advantage of these tasks for deposit of materials in repositories.

In addition to the proposal submitted, we have collected feedback on the Annual Report System that was developed for last year’s reporting cycle, and requested changes to the system in preparation for the current reporting cycle. Updates include an auto-complete function for filling out names of people and clarifications of wording and examples throughout the system.

Future Directions

- We will continue to add items to the repository
- We will continue to assist authors in the depositing process
- We will begin work on developing the CENS approach to the larger UCLA community

DATA 07 CENS Research Testbeds

CENS maintains a number of testbeds to enable and encourage various degrees of “in lab” to “real world” deployments using a spectrum of sensor network components, ranging from sensor hardware to network protocols. Our stationary testbeds include both indoor and outdoor deployments; outdoor usage is primarily in terrestrial settings, both above- and below-ground and is typically in medium- to-dense foliage environments. Our largest stationary testbed is currently deployed in the Andes. These stationary testbeds are augmented by a portable sensor platform that has been used in locations ranging from sea level to 15,000 feet in California and Central America and mobile device-based testbeds with over 500 devices to support urban and participatory sensing.

CENS Mobile Personal Sensing Testbed

With over 500 mobile devices, the CENS Mobile Personal Sensing Testbed supports several simultaneous real world pilot tests. The devices include nearly 500 mobile smart phones from Nokia, Samsung, and other manufacturers, running Symbian, Windows, and Android operating systems. In addition, a pool of 75 bluetooth GPS units can interface to mobile phones, and 20 GPS loggers can be used as standalone data collection devices. CENS maintains a variable number of cellular voice and data service plans as part of the testbed. Contact: Betta Dawson (betta@cens.ucla.edu).

CENS Central CA River Testbed

Based along a well-characterized reach of the Merced River at a USGS research site, the CENS river testbed includes shallow and deep (to 50 m) groundwater wells along a 1 km flow path from orchards, through row crops and a riparian zone to the Merced River. This testbed can be used in conjunction with the NIMS testbed (see below) to establish a system capable of characterizing both the river and the groundwater inputs and sinks. Contact: Tom Harmon (tharmon@ucmerced.edu).

CENS USC Coastal Robotic and Buoy Testbed

The NAMOS team at USC maintains a coastal monitoring testbed comprising static buoys and robotic boats. The buoys monitor the physical and chemical parameters of the water and communicate with the boat which can navigate according to the conditions observed by the buoys. In addition, the group uses two Slocum underwater gliders in coastal waters to monitor and communicate conditions that give rise to harmful algal blooms. Contact: Garuav Sukhatme (gaurav@usc.edu).

UC James Reserve Cyclops testbed

The 29-acre UC James Reserve, located about three hours from UCLA in the San Jacinto Mountains near Palm Springs, hosts several testbeds. The CENS Cyclops team has designed and deployed 25 imaging and environmental nodes to monitor active nest boxes spread around the James Reserve. These CENS-developed nodes—a synergistic combination of Cyclops boards and Crossbow nodes—record images and data every 8 to 10 minutes around the clock during the nesting season and incorporate CENS-designed software protocols. Contact: John Hicks (jhicks@cens.ucla.edu).

UC James Reserve AMARRS testbed

Underground soil sensing at JR incorporates a set of COTS data loggers augmented by nodes that produce real-time data streams to aid in monitoring both sensor health and real-time data analysis. This novel hybrid approach also leverages the ESS software. Contact: John Hicks (jhicks@cens.ucla.edu).

UC James Reserve NIMS testbed

The Networked Info-Mechanical Systems project has constructed a cable-based testbed at JR: NIMS-1 is a 35m transect spanning a creek bed and adjacent hillsides, with a controllable robot hosting a range of sensors (micro-climate, imaging, etc.). Contact: Eric Graham (egraham@cens.ucla.edu).

NIMS-RD portable testbed

A NIMS platform specifically designed for “rapid deployments” (in contrast to the permanent testbed installation at James Reserve) has been used in a remarkable variety of settings, measuring both terrestrial and aquatic phenomena. The NIMS-RD testbed has visited the 15,000 foot White Mountains of California to study plant growth and solar radiation effects, has been deployed many times on the Merced and San Joaquin rivers to study confluence patterns of aquatic contaminants, has been taken to the La Selva Reserve in Costa Rica to examine plant growth at the edge of the tropical forest, has been deployed across Lake Fulmor (a small lake adjacent to the James Reserve), and in Argentina to study lake contaminants. Contact: Bill Kaiser (kaiser@ee.ucla.edu).

CENS seismic testbed

This testbed is composed of 50 solar-powered wireless seismic nodes plus 15 wireless repeater stations, and is currently deployed in southern Peru where it spans the Andes with an ad hoc WiFi network. The testbed serves as a platform for experimentation with disruption tolerant network protocols while collecting seismic data studied in collaboration between the UCLA Earth and Space Sciences department and the CalTech Tectonic Observatory. It is relocated every 2-3 years: from 2005-2007 it was deployed in central Mexico; in 2008 it was moved to Peru; one half of the current transect will be relocated about 200 miles north in mid 2010. Contact: Richard Guy (rguy@cens.ucla.edu).