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)