1 Overview

1.1 Motivation

The NSF Next Generation Cybertools program has the ambitious goal of producing technologies that "not only change ways in which social and behavioral scientists research the behavior of organizations and individuals, but also serve sciences more broadly." This goal is particularly salient because the increased automation and "digitization" of work creates a sea of information about organizations and their processes. The availability of data creates the potential to revolutionize the way we understand, design, and manage organizations. To gain insight from this sea of data (rather than being drowned by it), we need ways to find patterns, interpret them and generalize appropriately.

In commercial organizations, opportunities to exploit improved mechanisms for qualitative and quantitative data exist in every core business process, such as new product development, customer support, supply chain management, and basic accounting. In addition to competitive pressures for process control and improvement, which date back to the early days of scientific management [57], commercial organizations are facing increased demands for compliance monitoring and internal controls [38]. Technologies, such as Enterprise Resource Planning systems, and continuous assurance auditing systems [73] create a virtual tidal wave of quantitative accounting data, but organizations lack effective ways to integrate the qualitative data needed to interpret it [38].

Many analogous opportunities exist in government and defense, as well. For example, military training and operations generates enormous amounts of detailed operational data that must be analyzed and interpreted [22]. Like commercial organizations, military operations include multiple, distributed participants, multiple hierarchical layers, and qualitative and quantitative data from many sources. Current technology for doing interpreting this data (e.g., Distributed Battlefield Exercise Simulation and Debriefing) focuses on one exercise at a time [47]. As with commercial organizations, the military faces significant challenges in gaining insights from qualitative and quantitative data generated by diverse sources [22, 47].

To frame our approach to cyberinfrastructure for organizational research (TestBed I), we begin by describing an organization with a host of interesting research opportunities and challenges directly related to this solicitation: the Defense Advanced Research Projects Agency (DARPA) High Productivity Computing Systems (HPCS) program [1].

The mission of the HPCS program involves the development of next generation, peta-scale high performance computing platforms for commercial availability by 2010. In a radical break with past high performance computing initiatives, the focus of this program is not just on the development of new and faster hardware. In addition, an explicit objective of this program is to decrease radically the cost and time required by organizations to perform their science and engineering activities that require these high performance computing environments. For example, the development of a new climate model might currently require a team of dozens of scientists and engineers several years to implement. Next generation HPC environments should simultaneously halve the size of the team and the time required to implement such a system. DARPA is currently funding research and development by IBM, Sun Microsystems, and Cray to better understand the hardware, software, and organizational requirements to achieve up to 10x productivity improvements.

Two of the principal investigators on this proposal have been associated with the HPCS program as academic researchers. This has given us insight into the enormous challenges associated with measuring, understanding, assessing, and improving organizational behavior in the largely unstudied domain of high performance computing system application development. While still in a very early stage, research by the vendors and affiliated researchers has begun to generate a body of quantitative and qualitative data concerning the behavior of developers and others in HPC organizations.

For example, pilot studies have been performed in a classroom setting with students developing simple high performance systems, resulting in quantitative data on the tools they used, the times at which they invoked the tools and the results, and properties (such as the size) of the software they produced [32]. Examples of qualitative data range from interviews with administrative staff of high performance computing centers to journals kept by professional developers as they work on HPC software [74].

The HPCS program and its organizations are confronting a variety of organizational research challenges directly related to the goals of the Next Generation CyberTools program.

First, the HPCS program is revealing the need for primary research on organizations using high performance computing environments. Basic questions need to be answered: How are high performance computing system applications developed and maintained? Where are the productivity bottlenecks? What are the organizational constraints on innovation in technology or methods? What is the most appropriate research methodology, or combination of methodologies, for gaining insight into these questions? This primary research will require the collection of substantial amounts of qualitative and quantitative data from a variety of contexts that must be disseminated to a broad range of users for a diversity of analyses.

Second, the answers to these basic questions must support the design of new technologies and organizational procedures that will yield an order of magnitude productivity improvement in high performance computing applications. This requires the operational definition and empirical validation of a productivity measure, generation of tools to collect the data necessary to calculate the productivity measure, and deployment of these tools in different computational environments and application domains.

Third, the HPCS program serves as an umbrella over many different types of organizations, generating substantial challenges regarding the publication and/or protection of information. The three HPCS vendor awardees, Sun, IBM, and Cray, are motivated to publish certain types of research results regarding productivity in order to (for example) influence the ultimate definition of the productivity measure used to evaluate their systems. On the other hand, each organization also generates research results that constitute proprietary information. The ultimate end-users of these systems (government and military laboratories, automobile companies, financial service institutions, etc.) form another set of organizations. The academic and corporate researchers form a third set of organizations. Collection and dissemination of qualitative and quantitative data amongst these organizations requires mechanisms for protection of privacy as well as proprietary trade secrets.

Fourth, the HPCS program is distributed geographically and involves a large number of constituent organizations and concurrent research activities. A major challenge to the program involves the requirement for alignment among the many approaches to qualitative and quantitative data gathering and research methods. An effective alignment will enable replication, in which data gathered to test a hypothesis at one site can be gathered in a similar manner at another site in order to see if the hypothesis is similarly supported. Alignment will also enable meta-analysis, in which data from multiple sites can be validly composed together into a larger dataset for the purpose of certain analyses.

We will return to the HPCS program in the Research Plan, where we will propose to deploy our cyberinfrastructure into it as part of a case study to evaluate our methods and technologies.

1.2 Cedar: Cyberinfrastructure for Empirical Data Analysis and Reuse

In this research, we propose to design, implement, and evaluate Cedar: a CyberInfrastructure for Empirical Data Analysis and Reuse, to satisfy the requirements for Testbed I. Cedar is intended to be an open source information infrastructure architecture coupled with a data management policy mechanism that supports scalable and collaborative, qualitative and quantitative organizational research data collection, analysis, dissemination, and archiving.

By *open source*, we mean not only that Cedar's source code will be released under a license that allows

access and modification by others, but also that we intend to create a community of developers willing and able to maintain and enhance the Cedar system beyond the period of this grant.

By *information infrastructure architecture*, we mean that Cedar will not be a monolithic system, but instead will specify a set of interfaces that allow integration and interoperability of tools for qualitative and quantitative data collection, analysis, and dissemination that will be developed both by us and by others.

By *data management policy mechanism*, we mean that Cedar will implement procedures that support context-sensitive publication, suppression, or perturbation of raw or processed qualitative or quantitative data, and support evolution in the policies applied to any specific data item over time. Appropriate data management policies should also generate incentives for data contribution and dissemination.

By *scalable and collaborative, qualitative and quantitative organizational research data*, we mean that Cedar will provide a federated network of peer-to-peer servers, creating scalability to thousands of concurrent data collection and analysis activities, and allowing analysis and annotation of data by many researchers across many institutions.

Finally, by *collection, analysis, dissemination, and archiving*, we mean that Cedar will support data management policies across the entire lifecycle of qualitative and quantitative data.

Cedar is an ambitious project that will require efficient and effective research and technology development in order to achieve its objectives during the grant period. At a high level, the project will focus on the following activities:

(1) *Infrastructure technology research and development.* Through the Hackystat Project, Principle Investigator (PI) Johnson has developed expertise in the development of open source collaborative systems for collection and analysis of quantitative data for software engineeering research and experimentation. The Hackystat system and experiences provide a base for extension into qualitative data collection and analysis, as well as to a peer-to-peer network of federated servers.

(2) *Research on and development of policies and procedures for data privacy and dissemination.* PI Basili is leading a task force of software researchers with experience in developing and maintaining software engineering empirical data repositories with the goal of articulating prior problems and proposing improvements for management of future repositories. We will leverage this initial research and incorporate related research in privacy policies and technologies for integration into the Cedar infrastructure.

(3) *Research on and development of models and mechanisms for representation and integration of qualitative and quantitative information.* PI Pentland and PI Feldman have carried out a variety of research on the theoretical underpinnings of qualitative and quantitative empirical data and its appropriate interpretation. Cedar will leverage these insights with technological infrastructure for collection, analysis, and dissemination of empirical data according to narrative and network theories for representation and analysis of qualitative and quantitative data.

(4) *Case study evaluation of Cedar.* The four PIs (Johnson, Basili, Pentland, Feldman) have substantial prior experience in the design and implementation of case studies across a variety of application domains and organizational types. To test the validity of Cedar, and to understand its strengths and limitations, we will perform a case study with selected organizations involved in the DARPA HPCS program.

2 Related Work

2.1 Qualitative and quantitative data and its integration

A primary requirement for Cedar is to support collection, analysis, and integration of qualitative and quantitative empirical data. To understand our approach to this requirement, it is useful to first introduce what we mean by qualitative and quantitative data and their interrelationship.

Concepts. By qualitative data, we mean text, images, and other materials that have symbolic meaning for some cultural group. Qualitative data comes in many different forms, from structured interviews, to surveys and questionnaires, to life history narratives, to full blown cultural ethnography. Qualitative data is often textual, but may include graphics, audio, video, or even clothing, architecture, and other cultural artifacts.

By quantitative data, we mean numbers: interval or ratio measures, including counts or frequencies of occurrence of objects or events, as well as the variable properties of those objects or events. For example, one might count the number of people on a team, or the number of tasks they perform per unit time. One might also measure their average tenure in their current jobs. In the domain of software engineering, typical quantitative data might include the number of lines of code (LOC) in a software module, or the number of modules in a system. Quantitative data always has an explicit or implicit time dimension: it quantifies something at a given point or interval in time. Quantitative data can be derived from qualitative data. For example, one might count the occurrences of a particular behavior recorded in a researcher's field notes.

Of course, numbers do not speak for themselves. Like qualitative data, they derive meaning from context. For example, is 5 defects per 1000 lines of code high or low? Is a $100,000 error in a financial statement "material" or not? While numbers might seem objective, qualitative data is often essential to making sense of quantitative data. The great strength of qualitative research is its ability to introduce context into the study of a particular phenomenon. From the beginning of the research process (research design, access to research sites, data gathering) to the end (analysis, writing and publication) qualitative research both necessitates and enables attention to context.

PI Feldman has carried out a variety of research focused on the question of how to define systems of meaning from qualitative and quantitative data [29, 30]. Her research illustrates how the analysis of qualitative data always involves at least two systems of meaning: that of the subjects being studied (the "participants", sometimes called "natives" or "insiders"), and that of the researchers, who have a theoretical framework. The participant perspective is referred to as "emic", while the researcher perspective is called "etic" [36]. Within an organization, there may be several different "insider" perspectives, as well.

People, including researchers, often make sense of the world and their place in it as a form of "narrative" [19, 33, 60, 52, 5, 75, 59]. Narrative provides context: it reveals what is significant to people about various practices, ideas, places, and/or symbols [76]. Narrative structure can form the basis for an analytical framework that connects the actions and events with the meaning(s) that these actions have for the people who take them [10]. With appropriate representational support, narrative structure appears promising as a means to integrate qualitative and quantitative data.

In summary, we believe that proper interpretation of qualitative and quantitative data requires context, that the data and context together form one or more systems of meaning, and that narrative structure forms an approach to integrated representation of qualitative and quantitative data. With these concepts in hand, we next review research related to technological support for collecting, analyzing, and integrating qualitative and quantitative data.

Qualitative data collection tools. The very nature of qualitative data limits the kinds of tool support for its collection. Diaries, field notes, interviews, and so forth can be readied for analysis with software for automated transcription or handwriting recognition. Questionnaires can be provided in an electronic form, such as over the internet. For example, Net-MR supports online surveys in 35 languages for multi-country data collection.

The state of the art in automated coding and analysis systems is advancing rapidly. For example, the Kansas Event Data System (KEDS) parses newspaper articles on political events in the Middle East and subjects to extracted event data to statistical analysis with the goal of predicting political change in the region [2]. In research for the National Gallery of the Spoken Word, researchers are using hidden markov models (HMM) for speaker-independent keyword recognition. By adding meta-data (codes) to raw data, these systems prepare qualitative data for subsequent analysis.

Qualitative analysis tools. Qualitative analysis tools are generally of two types: coding support tools and text mining tools. Coding support tools, such as ATLAS.ti, MAXqda, N6, NVivo, ETHNO, and Qualrus, allow researchers to annotate textual or video data with the goal of identifying the meanings implicit or explicit in the data. These tools help to speed analysis without disconnecting the data from the context more than is necessary. In general, current qualitatitive analysis tools are designed to support a small team of researchers working with a relatively small set of data (e.g., a set of interviews or fieldnotes). They do not support integration of quantitative data, support for large-scale distribution, analysis, or dissemination, or flexible privacy protection mechanisms.

Quantitative data collection tools. Collection of quantitative data is generally more amenable to automated support than qualitative data. Through the NSF sponsored Hackystat Project, PI Johnson has been investigating "sensor-based" approaches to automated, unobtrusive collection of quantitative data regarding software development products and processes. Hackystat is based upon his prior research on the Personal Software Process, which identified both logistical and quality problems with manual collection and analysis of quantitative software engineering data [41, 43].

Hackystat sensors are small, custom software "plug-ins" to developer tools such as editors, testing tools, configuration management systems, build tools, and so forth. Once a sensor is installed, it monitors the use of the tool and automatically sends these raw process or product data to a web server where further analyses can be performed. Examples of sensor data include: activities (compilation, file editing, etc.) within an editor, invocation of unit tests and their results, size/complexity of the system (in LOC, methods, classes, operators), configuration management events (such as commits and lines added/deleted), test case coverage, software review data (time spent doing review, issues generated, etc.) and defect management data. Specialized configurations of Hackystat have been developed for a variety of contexts, including classroom settings to support Java development [44], at the Jet Propulsion Laboratory to analyze workflow [42], and in case studies of high performance computing [46]. Figure 1 illustrates the Daily Diary, one of the perspectives provided by Hackystat for viewing sensor data. This Daily Diary instance is configured to show the commands entered by this developer into a shell and the most actively edited file (if any) during each five minute interval.

The automated nature of quantitative data collection in Hackystat allows any single installation to scale to dozens or hundreds of users. For example, the public Hackystat server maintained at the University of Hawaii contains accounts for over three hundred users and over 10,000 developer days of data. However, Hackystat does not support qualitative data collection, and implements a static privacy policy.

Quantitative analysis tools. Excellent tools exist for the analysis of quantitative data is through variance-based models, such as regression, structural equation modeling, event history, and so on. In the familiar regression framework, we create a model of the form Y = f(X), which posits a functional relationship between a set of antecedents (x1, x2, x3, ... xn) to a set of outcomes (y1, y2, y3, ... ym). Since all the variables can be expressed with numbers, we can use covariance-based methods to estimate the relationships and test their statistical significance. In software engineering, for example, the COCOMO cost model predicts outcomes concerning the cost and time associated with a software project given antecedents characterizing the system to be built and the resources available for its construction [17].

Integrating qualitative and quantitative data. In variance models, causal mechanisms are usually implicit [5, 50, 34]. Our quantitative methods allow us to demonstrate that Y = f(X), but documenting the chain of events that connects X and Y requires qualitative (narrative) analysis [5, 34, 23, 37]. Abbott has argued that significant new insights can be gained by using narrative models to investigate the patterns of events or actions that connect important antecedents and outcomes [4, 5, 6]. This insight forms the basis for our proposed approach to the integration of qualitative and quantitative data. PI Johnson has demonstrated this approach in the analysis of quantitative data in Hackystat called "Software Project Telemetry" [45]. Instead of building a predictive model to connect antecedents to outcomes, telemetry-based analyses focus on in-

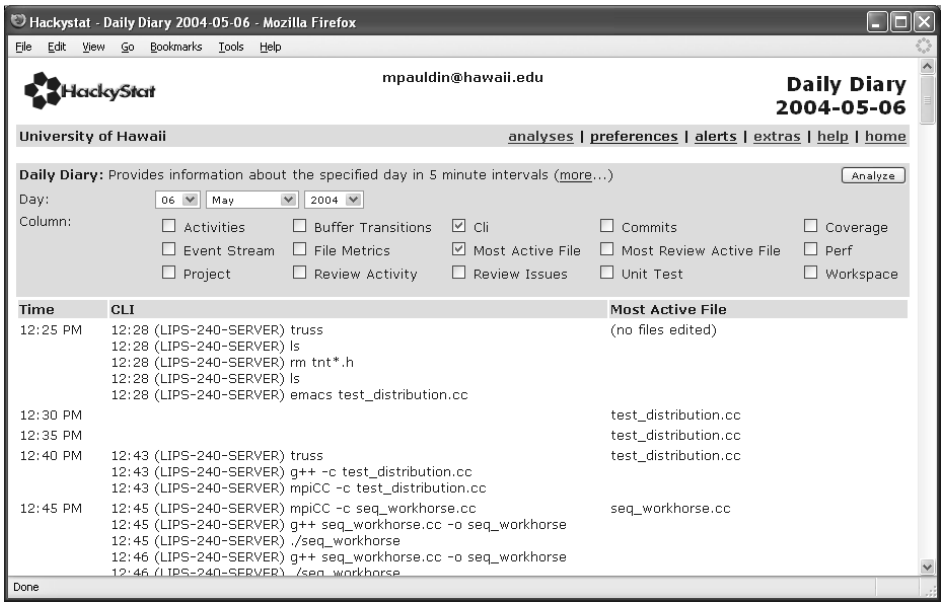

Figure 1. Hackystat Daily Diary, illustrating an event stream (shell commands) and quantitative data (most actively edited file during a five minute interval)

process monitoring of data streams and their relative change over time. For example, if test coverage values begin to trend downward and at the same time defect reports begin to trend upward, the project managers might hypothesize that a deterioration in testing quality is making an observable impact on software quality. Ongoing research is evaluating Software Project Telemetry for decision-making in the context of a daily build process. Figure 2 shows an example of software project telemetry which charts the relative growth of serial and parallel lines of code in a high performance computing application over a 12 month period.

Modes of integration. Our approach supports three basic modes of integrating qualitative and quantative data, each of which has significant body of related work in the social and organizational sciences.

(1) Counting and aggregating. Given a stream of qualitative (or quantitative) data stored in CEDAR, such as events, they can be counted and aggregated in various ways. This is a familiar analytical technique, and we do not see it as a significant research issue.

(2) Identification of causal patterns. Because CEDAR will store sequences (streams) of events, it should support efforts to identify patterns of events and determine the chain of events that connect antecedents and consequences. Similarly, optimal string matching [3, 4, 7, 6, 63] has been applied to a variety of organizational situations. PI Pentland has applied string matching to actions in a work process, using algorithms developed in molecular biology for the analysis of genetic sequences to compare and cluster sequential patterns [57]. Event structure analysis (ESA) [37, 23, 34, 35, 70, 71, 72] provides another methodology for interpreting events captured in ethnographic fieldnotes in terms of coherent patterns.

(3) Contextualization and interpretation. As mentioned above, traditional qualitative analysis requires putting data in context. The representation we propose to develop for Cedar (discussed below) will allow users to analyze qualitative and quantitative data using network techniques. Network representations provide a powerful means of contextualizing and interpreting qualitative data, as in semantic networks [67, 21], "cause maps" [54, 18] and "networks of action" [26, 56, 8]. PI Pentland has investigated the use of network models to represent interaction processes, and has developed a conceptual framework for the use of narrative data in the analysis of organizational processes [58, 56].

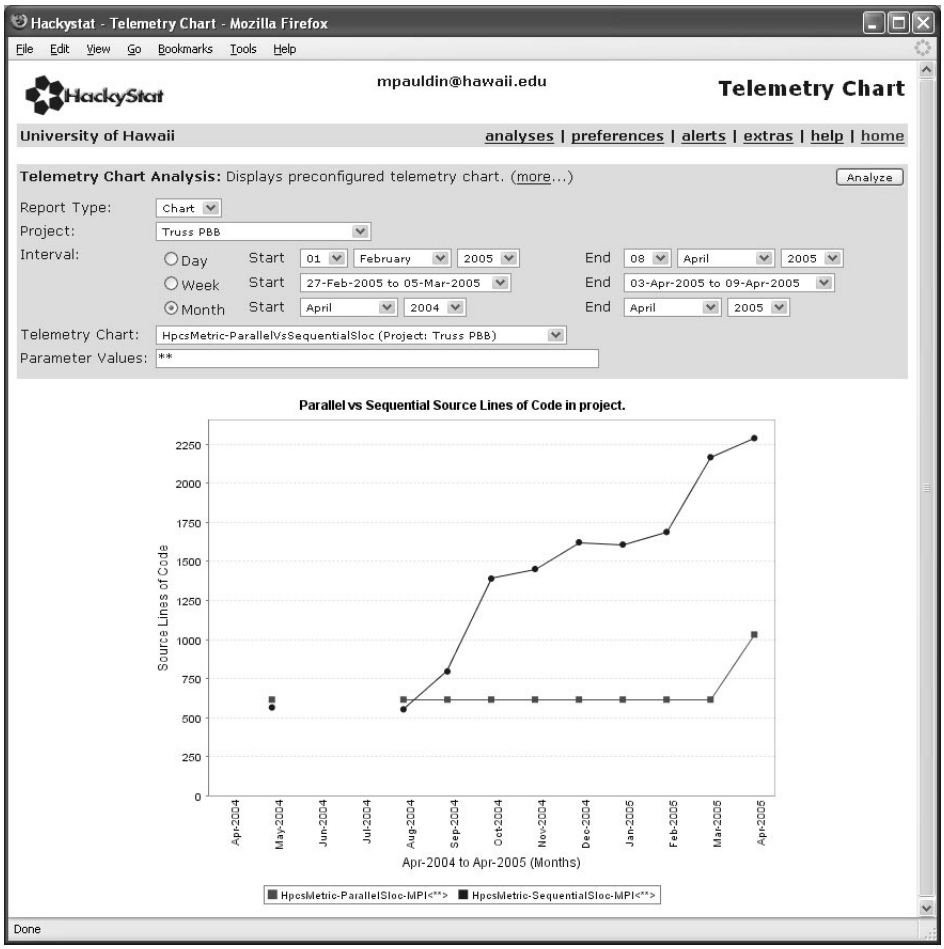

Figure 2. Example Hackystat Telemetry, showing trends in the amount of parallel and serial code in a high performance computing application over 12 months

Research Issues. Prior research indicates the promise of event-based analysis of qualitative and quantitative data, both separately and together. However, the research and technological innovations to date has been fragmented, both by discipline and by application area. Qualitative analysis techniques cannot scale to analysis of dozens or hundreds of subjects supported by Hackystat. On the other hand, the interpretation of event-based data collected by Hackystat could be improved by the narrative and network modeling techniques available from social science research. Finally, all of the research and technology suffers from an inability to interoperate with each other.

2.2 Infrastructure and experimental data repositories

An important component of empirical study is the generation of data, artifacts, and the use of experimental testbeds of various kinds. Infrastructure, such as online data repositories, makes this information available to others in the community to enable experimental replication, meta-analysis, and other applications. In this section, we present some examples of prior work on infrastructure and experimental data repositories for empirical research, followed a summary of the common themes.

NASA SEL database. Since its inception in 1976, PI Basili has been affiliated with the NASA Software Engineering Laboratory (SEL). The original mission of SEL was to study software development for Ground Support Software at NASA/GSFC with the goal of improving the quality of the software developed [14]. The lab has collected a variety of case study data including: resource usage; changes and defects; process, project, and product characteristics; and process conformance on several hundred projects. The artifacts developed are production ground support systems for NASA missions. Models of cost, schedule, and quality were built using this data. For years, this data was given away freely and made available through two repositories, one at NASA and one at the Rome Air Development Center.

Issues with the SEL experimental data repository experience include: misuse and misinterpretation of the data due to a lack of context information, leading to publications with questionable results; secondary data analyses were not known or shown to SEL; overhead of organizing, publishing, and supporting the repository; lack of support for feedback and meta-analysis. repository. Late in the life cycle of the project, it was recommended that anyone who wanted to use the data must first spend some time at the SEL in order to acquire an understanding of the nature of the data and its appropriate use.

CeBASE. More recently, PI Basili (along with Barry Boehm of USC) developed the Center for Empirically-based Software Engineering (CeBASE). CeBASE is an NSF sponsored project with the role of acting as a repository of "experience" on the effects of applying a variety of techniques, methods and life-cycle models to software development. Initial focus areas are defect detection methods [15], COTS development [11], and Agile methods. Data available in CeBase is currently qualitative in nature and is manually organized by the website maintainers. While CeBase contributors become publically affiliated with the project, as with the SEL repository, the data is made publically available from the website and there is no monitoring or control over dissemination.

Although CeBase is a much younger repository than SEL, many of the same issues regarding appropriate use of the data are expected to apply: how to ensure that secondary analyses of data is done appropriately, and how to support the ongoing overhead of repository maintenance and support.

Hackystat. As noted in the previous section, PI Johnson has been leading the Hackystat Project, which supports automated collection and analysis of quantitative software engineering data and which produces an information repository suited to empirical experimentation. The Hackystat data repository contrasts in an interesting way with both the SEL or CeBase repositories. While SEL and CeBase repositories are populated by data from completed projects or case studies, the Hackystat data repository collects data incrementally, in real-time, as a project progresses. While access to the information in the SEL and CeBase repositories is unlimited and uncontrolled, access to Hackystat data is strictly controlled: only the data owners or members of the project can access the data and use it for empirical analyses of their projects. For this reason, a central issue for the Hackystat data repository is effectively the opposite confronted by SEL and CeBase: how to make the Hackystat data public, and what form should that public data take?

Other experimental data repositories. Hackystat, SEL, and CeBase illustrate some of the experimental data repositories with which the PIs have had direct experience, but other repositories exist. For example, TheDataWeb is an online information repository for demographic, economic, environmental, health, and other datasets. Developed through a collaboration between the U.S. Census Bureau and the Centers for Disease Control, TheDataWeb provides unified access to data housed in different systems in 16 different federal agencies. Example datasets include American Housing Survey, the Behavioral Risk Factor Surveillance System, the Consumer Expenditure Survey, the Current Population Survey, and the National Center for Health Statistics Mortality Survey. From an implementation perspective, TheDataWeb consists of two parts: a "DataWeb Servlet System", which providers of data can use to make their datasets accessable to users of the TheDataWeb, and the "DataFerrett", a client-side application that enables users to browse and query these datasets and extract data from them.

Research Issues. A central theme from the research to date on experimental data repositories is the

importance of flexible control over access and dissemination of experimental data. In the case of SEL, too much access has led to inappropriate use of the experimental data and publication of questionable results. In the case of Hackystat, the fine-gained nature of the data and presence of personally identifying information has resulted in a policy of too little access, such that this rich source of experimental data is not available for secondary analysis, meta-analysis, replication, and so forth. Finally, lack of communication mechanisms between these repositories limits the possible insights to be gained by data aggregation and meta-analysis.

2.3 Privacy policies

As the prior section demonstrates, an important unsolved question in experimental data repositories is how to control the access and dissemination of the stored data. Such privacy policies must take into account a variety of issues.

The first, and most obvious issue, is to protect the privacy of the participants providing data to the repository. Singer provides an excellent overview of the ethical issues involved with data collection and storage of human subject data from empirical studies [66]. She analyzes the literature on research ethics and identifies four general principles for ethical empirical experimentation: informed consent, scientific value, beneficence, and confidentiality, and illustrates their application or misapplication in a variety of experimental contexts.

Singer makes the important point that compromising these guidelines not only risks the exploitation of individuals associated with the study, it also risks harming the study, the subjects themselves, or the organization under study. For example, missing or insufficient application of informed consent or confidentiality could lead to subjects not cooperating or providing incorrect data. In the case of their managers, it could lead to loss of access to study sites or loss of funding. Both of these harm the study. Inappropriate access to empirical data could reveal data that could be used to damage an individual's professional reputation, harming their career. Finally, inappropriate interpretation of data could lead to negative consequences for the organization in which the study took place. Austen provides an additional perspective on this kind of institutional harm called "measurement dysfunction" [9]. Such "secondary" use of data and its privacy implications have been investigated in a health care setting [51].

Singer applies the principles of informed consent, scientific value, beneficence, and confidentiality somewhat narrowly to the subjects and organizations who provide the data to the repository. Our prior involvement with public data repositories reveals the need to expand these principles in our proposed research. For example, the SEL experience demonstrates the real risk that public repositories of empirical data can lead to misinterpretation of the extracted data due to inadequate contextual information (i.e. meta-data). Our Hackystat experience demonstrates that "boilerplate" application of standard informed consent practices in an educational setting leads to a repository whose data cannot be effectively "freed" for purposes such as third-party meta-analysis. The presence of an online, publically available data repository creates the need for privacy policies that enable control not only over the type and level of contributions to the database by subjects, but also on the type and level of data extracted from the database by users. Indeed, for a long-lived data repository, such control might evolve over time: data contributed regarding an organizational project might have quite restricted access at the time it is initially contributed due to risks of leaking proprietary data. Five or ten years later, a more relaxed level of control over access might be possible, rendering new types of analysis possible.

Some prior work has been done on the technology of privacy specification and implementation. For example, the Platform for Privacy Preferences project has developed a way for users to specify and change their privacy preferences with respect to their interactions over the Internet [25], though other research has indicated that what people specify as their preferences may not reflect their actual behavior [68]. Research is also available on sanitizing or anonymizing private data for the purpose of data mining [61, 39]. These

approaches typically involve perturbation of data values, generalization/abstraction, or suppression.

Research issues. To build an effective cyberinfrastructure for qualititative and quantitative data collection and analysis, we must be aligned with the four principles of research ethics, expand them to address control over collection and dissemination and its evolution over time, and leverage the technology that is currently available.

2.4 Results from prior NSF research

Award number:	CCF02-34568
Program:	Highly Dependable Computing and Communication Systems Research
Amount:	$638,000
Period of support:	September 2002 to September 2006
Title of Project:	Supporting development of highly dependable software through continuous, automated, in-process, and individualized software measurement validation
Principal Investigator:	Philip M. Johnson
Selected Publications:	[46, 55, 45, 44, 43, 42, 49, 28, 48]

The general objective of this research project is to design, implement, and validate software measures within a development infrastructure that supports the development of highly dependable software systems. Contributions of this research project include: (a) development of a specialized configuration of Hackystat to automatically acquire build and workflow data from the configuration management system for the Mission Data System (MDS) project at Jet Propulsion Laboratory; (b) development of analyses over MDS build and workflow data to support identification of potential bottlenecks and process validation; (c) identification of previous unknown variation within the MDS development process; (d) development of a generalized approach to in-process, continuous measurement validation called “Software Project Telemetry”, (e) substantial enhancements to the open source Hackystat framework, improving its generality and usability; (f) development of undergraduate and graduate software engineering curriculum involving the use of Hackystat for automated software engineering metrics collection and analysis; (g) support for 3 Ph.D., 6 M.S., and 3 B.S. degree students.

Award number:	CCR-0086078
Program:	Information Technology Research
Amount:	$2,400,000
Period of support:	September 2000 to September 2003
Title of Project:	ITR: Collaborative Research Proposal for a National Center for Empirical Software Engineering Research
Principal Investigator:	Victor Basili, Barry Boehm
Selected Publications:	[12, 13, 16, 40, 53, 62, 64, 65]

The CeBase research activities have allowed the current research team to build significant expertise in the areas of software engineering decision support, defect analysis, and empirical study that are vital for the proposed work. Contributions of this research include: (a) interaction with an official ”affiliates list” of over 21 university, industry, and other research organizations; (b) Three tutorials in empirical research methods and empirical results; (c) 11 end-user forums for CeBase end-users; (d) Development of a publicly-available repository, www.cebase.org, containing research tools, reusable artifacts and documents, supporting data, and results. (e) Development and public release of tools (such as eWorkshop) for use by the empirical research community; (f) 9 books and book chapters, 46 refereed journal publications; 57 refereed conference and workshop publications; (g) support for 10 Ph.D., 6 M.S., and several B.S. degree students.

3 Research Plan

We begin by presenting our research plan as three high-level initiatives: the design, implementation, and evaluation of the Cedar infrastructure. We then present a more low-level view of the plan in terms of tasks, milestones, coordination activities, outreach, and dissemination.

3.1 Design of Cedar

Section 1.2 summarized the high level requirements for the Cedar system: an open source information infrastructure architecture coupled with a data management policy mechanism that supports scalable and collaborative, qualitative and quantitative organizational research data collection, analysis, dissemination, and archiving. The risk level associated with each requirement varies according to whether it necessitates advances in the state of the art. A primary goal of the Cedar design phase is to focus on the high risk requirements and perform the research necessary to reduce the level of risk associated with them. In this project, risk assessment, risk reduction, and system design is an ongoing, iterative activity. We have identified three design issues for initial risk reduction: privacy policies, repository data management, and representational support for qualitative and quantitative data integration.

Design of privacy policies. There is an inherent tension between privacy and utility with respect to empirical data. The more you know about the data and the context under which it was collected, the more likely you are to assign a meaning or interpretation to the data that aids in understanding and/or decision-making. At the same time, the more you know about the data, the less privacy exists with respect to the individuals and organizations associated with it. We do not expect to eliminate this tension in the design of Cedar, but rather to leverage our own prior experience and other research to invent better mechanisms to manage this tension between privacy and utility. For example, traditional empirical data privacy policies are neither context-sensitive nor time-dependent: a subject signs a consent form that specifies a single type of access by a single type of researcher which never changes. A cyberinfrastructure for empirical data enables a more flexible approach in which a subject could specify different levels of privacy for different groups of people. Furthermore, the desired privacy could evolve over time: an organization might require high levels of privacy for data associated with a product under current development, but might be willing to loosen privacy levels a decade later after the product has been retired. PI Johnson will lead this design effort.

Design of repository data management. Prior experience by both PI Basili and PI Johnson with public empirical data repositories indicate that "if you build it, they will come." However, many important issues remain: how can we help ensure that "they" use the data appropriately? How can we create appropriate incentives so that "they" want to contribute new data as well as extract insight from the old? How can we create a self-sufficient repository that supports the ongoing need for hardware, software, and technical support resources? PI Basili will lead this design effort as a natural extension of his ongoing leadership role in the software engineering task force on data repositories.

Design of narrative and network representation for integration of qualitative and quantitative data. Prior research by PI Pentland and PI Feldman indicate that narrative and network representations show great promise as a means to integrate qualitative and quantitative data, providing the context and etic/emic perspectives necessary to derive meaning from data. For example, both qualitative and quantitative data in Cedar could be indexed using categories from narrative analysis, such as Burke's grammar of motives (actor, act, scene, agency, purpose) or Fillmore's case grammar [20, 31]. The choice of exactly which dimension to include, and the extent to which customization is allowed, are important research questions. For example, we can add additional elements to the structure, such as "input" and "output", so that we could support generic process representations, such as the Process Specification Language (PSL). Questionnaires, surveys, spatial coordinates or other properties (e.g., generic keywords) could also be added when necessary, without loss of

generality for the basic indexing mechanism. PI Pentland has data on customer service processes in Citibank that can be used to support initial design and risk reduction. PI Pentland and PI Feldman will lead this design effort.

3.2 Implementation of Cedar

The Cedar implementation, and the process we use to achieve it, will be modeled in many ways on the Hackystat Project. Hackystat already embodies many of the attributes we need for Cedar, and we can leverage our experiences with the Hackystat development process and the software that has resulted to jump-start Cedar development. Hackystat has gone through five major architectural revisions since its inception, as we worked through issues related to scalability, extensibility, and configurability, and currently consists of approximately 100,000 lines of code. The current architecture enables us to implement both a generic infrastructure for collection and analysis of quantitative software engineering data, as well as a set of "configurations", or enhancements to this generic infrastructure that customize the data collected and the way it is analyzed to the needs of a specific situation. The Hackystat project is self-instrumented, and we use measurements of our own process and products to maintain a balance between quality assurance and enhancement activities.

The requirements for Cedar will require three major extensions to Hackystat: a federated peer-to-peer network of servers; support for qualitative data collection and storage; and an enhanced client-side user interface.

Federated servers. Hackystat implements a fairly standard client-server architecture: sensors collect data from client systems and sends it to a Hackystat server for storage and analysis. We maintain a public Hackystat server to which anyone can send data, though some organizations prefer to install and maintain their own server so that data about their processes and products remains internal. Cedar will extend this basic paradigm by allowing servers to establish communication with each other, creating a federated, peer-to-peer network of Cedar servers. This architecture will have an interesting impact on the design of privacy policies: it seems likely that a user will wish to establish both a "local" privacy policy (i.e. for how their data is protected in the server where it is physically located) and a "global" privacy policy (i.e. for how their data might be disseminated upon request to other servers.)

Qualitative data collection and storage. Hackystat does not currently collect or store qualitative data. Cedar will extend support to collection and storage of many, but not all, forms of qualitative data. Cedar will not, for example, provide the ability for users to upload a digital version of a feature film and store it along with indexes into various scenes of interest. Such kinds of qualitative data must be represented in Cedar indirectly: instead of storing the actual file containing the feature film, Cedar might store, for example, an URL to a location on the internet containing the film, along with information describing how to access the scenes of interest using some appropriate viewer.

User interface. Hackystat's user interface is web-based, consisting of HTML forms for entry of information and static tabular or chart data as the results of analyses. The advantage of a web-based interface is that users with only a browser can access and manipulate Hackystat services. The disadvantage is the constraints that HTML places on the way data is entered, displayed, and manipulated. For Cedar, we will design and implement a more sophisticated client-side application called CedarView that will allow display, entry, and analysis of both qualitative and quantitative information. CedarView will be inspired by multi-track editors for music, such as Apple's GarageBand. Upon execution, CedarView will connect to one or more Cedar servers, download the appropriate data, and display it as a series of "tracks" organized along a timeline. It will allow the user to "zoom in" or "zoom out" of the chosen data streams, and "cut and paste" data streams from one timeline to another. It will allow annotation of timelines with additional information, such as for encoding episodes with classifiers. Finally, CedarView will be extensible through a plug-in architecture to

support processing of the raw data in various ways. For example, one plug-in might produce a timed markov model, while another might produce a social network representation.

3.3 Evaluation of Cedar

Evaluation of Cedar will be an ongoing process throughout the project, where the artifacts to be evaluated and the approach to be used will depend upon the stage of development of the system. For example, early in the project, we will seek community review and evaluation of design documents regarding our approach to privacy policies, repository data management, and the representation and integration of qualititative and quantitative data. To facilitate this evaluation, we will form an advisory board from a variety of academic and industrial disciplines to ensure broad perspectives and feedback on our approach.

As elements of the Cedar infrastructure come online, we will begin a series of evaluative case studies to assess how well the infrastructure is fulfilling its requirements. We will begin with classroom use to assess basic functionality and usability. After this use indicates sufficient stability and functionality, we plan to incrementally deploy Cedar into the High Productivity Computing Systems organization, as described in Section 1.1.

Case studies in the HPCS organization will allow us to "stress-test" virtually all of functionality intended for Cedar. As a distributed organization, use of Cedar by HPCS will naturally lead to a set of intercommunicating servers. The relationships between the various organizations will test the ability of our privacy policies to enable sharing of data while providing adequate protection to the subjects and organizations who generated it. The diversity of qualitative and quantitative data will test the ability of Cedar to represent this information and make useful connections between them. Finally, the national importance of the HPCS program and the level of commercial and government investment in it provides natural incentives for long-term resources for management of the Cedar repository, and tests the abilities of our management policies to exploit those potential resources.

Use of Cedar in the HPCS domain will provide a body of experiences, data, and technology transfer insights that we will exploit to gain insight into the requirements for broader outreach and dissemination of this technology to the scientific community. We anticipate forming a "Cedar Consortium" of academic and commercial organizations, along with a yearly users group meeting to share experiences and develop plans for future growth. The Apache and Eclipse communities provide models for how the various legal, organizational, and development issues can be resolved to form a vibrant community of users and developers.

3.4 Coordination plan, timeline, outreach, and dissemination

Figure 3 outlines the major tasks, lead PI(s), and milestones we have planned for this project. Although all PIs will be in close communication and involved with all aspects of the project, we believe that some decoupling, particularly in the initial phase of the project, will enable us to make progress more quickly.

In the initial year, the primary tasks will be to design privacy policies, narrative/network representations, and data repository management mechanisms; and implement the architectural framework for Cedar. We will also perform some "pre-pilot" case study work to validate our current requirements and gather additional ones for deployment of Cedar into the HPCS organization. At the end of this initial year, we will make a public release of our framework along with documents specifying the results of our design activities.

In the second year, we will implement the designs developed during the first year, and begin classroom case studies with Cedar which will result in curriculum materials. By half way through the funding period, at the beginning of the third year, we plan to have a first release of the fully functional Cedar cyberinfrastructure.

While we plan to continuously refine and improve the system for the remainder of the grant period, this process will be driven by more comprehensive evaluation activities during the second half of the grant

Task	Lead PI(s)	FY 2005	FY 2006	FY 2007	FY 2008
Data Privacy Design	PJ				
Narrative/Network Representation Design	BP, MF				
Repository Data Management Design	VB				
Cedar architecture implementation	PJ				
First public release of Cedar Framework			X		
Data Privacy implementation	PJ, VB				
Narrative/Network analysis implementation	PJ, BP				
Cedar evaluation pilot studies	VB, MF				
Cedar curriculum development	MF, BP				
Cedar release with privacy/narrative/network				X	
Cedar implementation enhancements	PJ, BP				
Cedar case study evaluation	PJ, VB, MF				
Pilot courses on Cedar-based experimentation	BP, MF, VB				
Cedar release with sample data				X	
Technology transfer, lessons learned	BP,MF,PJ,VB				
Wide-scale, federated servers in public use					X

Figure 3. Work breakdown structure and milestones

period, including pilot courses, case studies in the HPCS organization, and external evaluation. In the final year, we will begin effort on technology transfer, developing documentation and training materials and conducting tutorials as necessary to help broaden usage of Cedar. By the end of the grant period, we plan for a federated network of at least 50 to 100 Cedar servers in active use, and that the success of this framework in the HPCS community has created interest and involvement in the Cedar Consortium by both academic and commercial organizations.

All of the PIs have extensive prior experience working in distributed "virtual" organizations, and we have learned how to be productive despite geographical separation. We will use a variety of synchronous and asynchronous mechanisms to facilitate communication and coordination among the Cedar research group. We plan to have a weekly teleconference meeting between the PIs and graduate students to discuss tasks and challenges. Our travel budgets include funds for on-site meetings twice a year for more intensive, face-to-face interaction where we can review progress and establish goals and milestones for the next six month period. Finally, we will provide a website similar to www.hackystat.org that will provide a portal for access to source code, documentation, tech reports, wiki collaboration, a public Cedar server, and so forth.

3.5 Cedar in action

To illustrate the benefits of Cedar, consider once again Figures 1 and 2, which illustrate some of quantitative data available about a high performance computing system development project. Effective interpretation and application of this experience raises many questions : Are the trends in serial and parallel code typical? Under what circumstances would a new development project produce the same size trends? What are the strengths and weaknesses of the chosen tool set (g++, mpiCC, etc.)? Answering these questions requires contextual, qualitative data, much of which is potentially available in other artifacts associated with this study (the developer's engineering logs, emails, and so forth).

One goal of Cedar is to provide an effective representation for tying the quantitative to the qualitative, and it accomplishes this by supporting the creation of a high-level abstract narrative, or "story" of this development project which incorporates both the quantitative numbers and the explanations for how the numbers came about. In this case, some of the relevent context is that the developer is a graduate student, was implementing his first MPI program, was more concerned with functional correctness than parallel speedup during this project, and encountered a major requirements change in February 2005, leading to the sudden perturbation in both parallel and serial code size. Such context is crucial for assigning meaning to

these numbers.

The narrative representation has another benefit beyond data integration: it also provides a way for a user to query the federated network of Cedar servers for similar "stories". For example, having constructed this narrative, one can then query the network to learn about other case studies involving, for example, MPI software development. The level of detail provided back in response to this query will depend upon the privacy policies in force related to each instance of the narrative. The incremental generation of a collection of narratives, related in various ways, creates a rich web of qualitative and quantitative data that provides context for each single narrative as part of a larger community of practice.

4 Conclusions

We believe the Cedar project has substantial intellectual merit: it brings together not only qualitative and quantitative data, but also researchers from multiple disciplines to synthesize their knowledge and capabilities to produce a system with unique capabilities. The four Principle Investigators in this project bring a diverse, but complementary set of skills regarding qualitative and quantitative data collection and analysis; software development; and empirical data repository management. Our application of narrative and network theories to integrating empirical data is both novel and promising, and our prior development of the Hackystat system enables us to jump-start the Cedar implementation.

We have designed the Cedar project with the intention that it be broadly applicable as a tool for collecting, analyzing, and disseminating qualitative and quantitative data. As the University of Hawaii is a university with 75% minority students in an EPSCOR state, this project will provide novel research opportunities to underrepresented groups. The development of curriculum materials, classroom evaluation, case studies in HPCS, and technology transfer through the Cedar Consortium will all result in enhancements to infrastructure for research and education.

We would like to conclude by noting that the ability to gather and access data of this sort brings with it a duty to develop ways of interpreting these data responsibly. We have already pointed out that existing data bases have been subject to misuse and misrepresentation. People who are marginal in society are particularly vulnerable to misinterpretation of their actions. Poor people, for instance, find a variety of ways of coping with the lack of money that may make them look like irresponsible parents or even criminals when neither may be the case [24, 27, 69].

An important goal of this research is to help people incorporate, rather than bypass, context so that interpretations are smarter. Cedar will enable data analysts to identify some of the multiple stories (or at least be aware of the multiple stories) and think about what questions they need to ask and who they need to ask them of in order to sort through which stories are more likely than others. We view this project as an opportunity not only to be more precise in the data we are gathering but also as a way to incorporate the intrinsic diversity of meanings in any set of actions. If we succeed, this project will allow us to make the complexity of life more accessible rather than to obscure it.