GarageLab/Datalla: Next Generation Information Infrastructure for Qualitative and Quantitative Understanding of Organizations

Philip M. Johnson
Department of Information & Computer Sciences
University of Hawaii

Victor R. Basili
Department of Computer Science
University of Maryland

Brian T. Pentland
Department of Accounting & Information Systems
Michigan State University

Martha S. Feldman
School of Social Ecology
University of California, Irvine

Contents

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."

To make the research issues related to this goal more concrete, and to frame our approach to addressing them, we begin this proposal by describing an organization with a host of interesting research opportunities and challenges: 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 radically decrease 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 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 [8]. 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 [20].

Initial analyses of the raw data have included formal models, such as Timed Markov Models fit to classroom data [19]. Other case study data has been used to generate semi-formal models, such as "telemetry" based analyses [11]. Still other kinds of data, such as the qualitative journal data, has been best suited to qualitative encoding techniques [20]. Research has also led to proposals for new ways to assess high performance productivity, such as Purpose-Based Benchmarks [9].

So far, dissemination of research data and results have been via HPCS program meetings [4] academic workshops [18, 10], and themed journal issues [13].

As the HPCS program builds momentum, a variety of organizational research challenges are appearing.

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?

Second, the answers to these basic question 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, of course, an operational definition of "productivity" that can be measured in both current and future environments. Interestingly, no such measure has yet been agreed upon by this community, even though its definition has profound implications for the evaluation of the technologies under development and the future processes and products of the end-user organizations.

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 openly 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. Each of these organizational layers have privacy issues related to the information they collect, manage, and disseminate to others.

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 ensuring alignment among the many approaches to qualitative and quantitative data gathering and research methods. A true "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.

Having set the stage, we now present the fundamental objective of this proposal, followed by an overview of the information infrastructure we will use to achieve it.

1.2 Objective

The objective of our proposed research is to produce an open source information infrastructure architecture and data management policies that support scalable, collaborative, distributed, integrated, qualitative and quantitative organizational research data collection, analysis, dissemination, and archiving.

By "open source", we mean not only source code released under a license that allows access and modification by others, but also the creation of a community of developers willing and able to maintain and enhance this infrastructure beyond the period of this grant.

By "information infrastructure architecture", we mean the creation of a software framework that allows integration and interoperability of tools developed by us and by others.

By "data management policies", we mean procedures and mechanisms that support context-sensitive publication or protection of raw or processed qualitative or quantitative data. The management policies will not only address privacy issues, but also "lifecycle" issues related to data repositories.

By "scalable, collaborative, distributed, integrated, qualitative and quantitative organizational research data", we mean an infrastructure that can support hundreds to thousands of concurrent data collection and analysis activities, allowing analysis and annotation of data by many researchers across many institutions, combining both qualitative and quantitative data.

Finally, by "collection, analysis, dissemination, and archiving", we mean an infrastucture that can support data management policies across the entire lifecycle of qualitative and quantitative data.

1.3 GarageLab/Datalla

Our information infrastructure consists of two fundamental components: GarageLab, a front-end system to support the display and analysis of qualitative and quantitative information, and Datalla, a back-end peer-to-peer network to support controlled dissemination of the collected data.

The basic function of GarageLab is to support display and analysis of data. First, it enables the researcher to visualize multiple streams of raw qualitative and quantitative data by organizing each as "tracks" along a timeline. Similar to multi-track editors for music (such as GarageBand [2]), GarageLab allows the user to "zoom in" or "zoom out" of the chosen data streams, and "cut and paste" data streams from one timeline to another. GarageLab will also allow annotation of timelines with additional information, such as for encoding episodes with classifiers. Finally, GarageLab will allow plug-ins 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 network representation.

Datalla, the back-end system, provides several services. First, each Datalla server provides storage for raw qualitative and quantitative organizational data. A user with an account on a Datalla server can login via GarageLab to access data on that Datalla server. Second, each Datalla server can receive qualitative and quantitative data from Datalla "sensors", which are small software programs that can be used to collect and send raw data to a server. Finally, each Datalla server can communicate with other Datalla servers, forming a peer-to-peer network. The kinds of data that can be communicated to other servers is controlled by the privacy policy in effect.

1.4 Research Approach

Achieving our objective using the GarageLab/Datalla infrastructure will require us to carry out the following research and development activities.

(1) *Infrastructure technology research and development.* Our prior experience with Hackystat [3] provides us with expertise in open source development of client-server systems for automated collection and analysis of quantitative data. We will leverage this experience in the development of the GarageLab and Datalla software infrastructure. Research challenges include successful application of the GarageBand multitrack metaphor to display and manipulation of qualitative and quantitative organizational data, and the development of suitable APIs to allow 'plug-ins' with appropriate access to internal data.

(2) *Research on and development of policies and procedures for data privacy and dissemination.* While the infrastructure can make context-dependent privacy policies possible, we must perform research to understand what appropriate privacy policies would be. Such policies will influence the design of publication/protection mechanisms within the infrastructure. Research challenges include an appropriate means to classify data with respect to its privacy policy, appropriate safeguards to prevent unauthorized dissemination, and evaluation mechanisms to determine the effectiveness of a privacy policy once in place.

(3) *Research on and development of models and mechanisms for integrating qualitative and quantitative information.* Our basic infrastructure can "integrate" qualitative and quantitative information only in a fairly superficial sense: the data can be stored in together in a repository, and the raw data can be displayed together along a timeline. True integration goes much deeper: how do the qualitative and quantitative data come together to tell us something new about the organization that we could not have known from either kind of data by itself? We will pursue network models [16] as one approach to this "deeper" integration. Research challenges include the dependencies between the integration models and the raw data required to successfully connect the two types of data.

(4) *Case study deployment in the high productivity computing systems organization.* To test the validity of our approach, we will perform a case study of infrastructure deployment with selected partners in the HPCS domain. Through this case study, we will evaluate how well we have accomplished each component of the objective stated above. Note that there is a "meta" level in this case study. At one level, HPCS researchers will be collecting, integrating, analyzing, disseminating, and archiving qualitative and quantitative data about high performance computing. At the meta level, we will be collecting qualitative and quantitative data about this usage of the GarageLab infrastructure and policies in order to evaluate our approach. Research challenges include ensuring that the technology is robust enough for use in a live environment, and gaining

buy-in from the case study participants necessary to evaluate the deployment effectively.

(5) *Case study deployment in an accounting organization.* While our access to the HPCS community and the breadth of challenges it poses forms a compelling case study environment, we do not want to produce an infrastructure and set of policies that are inadvertantly specific to this domain. To maintain the generality of our infrastructure, we will also carry out a case study in the domain of accounting. Like (3), this will also involve a “basic” and “meta” level of data collection and analysis. Research challenges include the same ones as in (3), plus the additional risk that the new domain may violate implicit assumptions made in the original formulation of the system and its mechanisms.

2 Related Work

To be written.

It should contain at least the following:

1. Overview of Hackystat, software project telemetry, sensor-based collection and analysis of quantitative data. (Johnson)
2. Overview of CeBase, ISERN, SEL, etc. and their recent efforts to develop policies and procedures regarding protection/publication/maintenance of quantitative data. (Basili)
3. Integration of qualitative and quantitative data, narratives, encoding, etc. (Pentland, Feldman)
4. Results from prior NSF-sponsored research (Johnson, Basili)

Note that our related work section should be written to highlight any of our prior work that might already address any of the Solicitation Goals summarized in Section 4.1.

3 Research Plan

To be written.

One approach would be to have subsections detailing the research to be taken for each of the four sections in the “Approach” section above:

1. Infrastructure Technology Development. (Johnson)
2. Development of policies and procedures for data privacy and dissemination. (Basili)
3. Case study deployment in the high productivity computing systems organization. (Johnson)
4. Case study deployment in an accounting organization. (Pentland/Feldman)

Again, our research plan must be constructed to show how we will address the Solicitation Goals summarized in Section 4.1.

4 Anticipated Contributions

To be written.

In this section and throughout the whole proposal, we need to both ensure that we are achieving the goals of this solicitation, as well as clearly addressing the standard NSF review criteria. To help us get there, the next two sections provide copies of text from the solicitation that we will need to keep in mind as we proceed.

4.1 Solicitation Goals

Some of the solicitation goals include:

1. The development of tools that facilitate the integration of qualitative and quantitative information from heterogeneous sources, multiple media, and/or multiple modes;

2. Investment in basic research that addresses the protection of the confidentiality of respondents in computerized, widely accessible databases; and

3. The development of incentives, standards and policies for collecting, storing, archiving, accessing, and publishing research results using organization-relevant information.

4. Testbed I. information collected on organizations from a variety of heterogeneous, independently developed data sources, such as administrative and survey data, temporal, spatial and image data or textual data. The goal is to free users from having to locate the data sources, interact with each data source in isolation, and manually combine data from multiple formats and multiple sources. This could be achieved through the creation of new and more accurate and efficient ways to collect, code and analyze qualitative information from case studies, and other sources, and to enable the linking of this information with repositories of quantitative data, while protecting fundamental privacy and confidentiality concerns. The research should be designed to show how appropriate cybertools can lead to multiple advances in the empirical understanding of how organizations emerge, develop, thrive or weaken.

5. Proposals must address the protection of data providers from identification, exploitation, and other misuses of personal or organizational information. Such misuses present a perpetual challenge to the melding of data and media of different types in a tool for widespread use. Proposals in response to this solicitation must show a sophisticated understanding of this sociotechnical problem and must propose to advance fundamental knowledge of effective privacy protections during the development of the analytical tools and in their later use by various research communities.

6. Proposals must demonstrate potential long-term sustainability, usability, and impact. This could be achieved for the organizational ”testbed”, for example, by documenting proposed collaboration with firms in an industry, attracting support from foundations or developing replicable incentive-compatible policies for collecting, storing, accessing, and disseminating data while continuing to utilize and advance relevant cybertechnology.

7. Unifying Data Models and System Descriptions: There is a need to develop stronger theoretical foundations for the representation and integration of information of various types from extant data models (e.g., temporal, spatial and image data, textual data, administrative and survey data) as well as the scientific literature into conceptually coherent views.

8. Reconciling heterogeneous formats schemas and ontologies: The fundamental problem in any data sharing application is that systems are heterogeneous in many different aspects, such as different ways of representing data and/or knowledge about the world, different representation mechanisms (e.g., relational databases, legacy systems, XML schemas, ontologies), different access methods and policies. In order to share data among heterogeneous sources, approaches to form a semantic mapping of their respective representations are needed to avoid manual intervention in each step of converting and merging data resources.

9. Web semantics: Data on the web needs to be defined and linked in a way that it can be used by machines not just for display purposes, but also for automation, integration and reuse of data across

various applications. Supported research topics will include frameworks for describing resources, methods of automating inferences about web data and resources, and the development of interoperable ontologies, mark up languages and representations for specific social, behavioral and other scientific domains.

10. Decentralized data-sharing: Traditional data integration systems use a centralized mediation approach, in which a centralized mediator, employing a mediated schema, accepts user queries and reformulates them over the schemas of the different sources. However, mediated schemas are often hard to agree upon, construct and maintain. For example, researchers conducting social and behavioral research share their experimental results with each other, but may do it in an ad hoc fashion. A similar scenario is found in data sharing among government agencies. Architectures and protocols that enable large-scale sharing of data with no central control are needed.

11. On-the-fly integration: Currently, data integration systems rely on relatively static configurations with a set of long-lived data sources. On-the-fly integration refers to scenarios where one wants to integrate data from a source immediately after discovering it. The challenge is to significantly reduce the time and skill needed to integrate data sources so that scientists can focus on domain problems instead of information technology challenges.

4.2 NSF Review Guidelines

The generic ones are:

1. What is the intellectual merit of the proposed activity? How important is the proposed activity to advancing knowledge and understanding within its own field or across different fields? How well qualified is the proposer (individual or team) to conduct the project? (If appropriate, the reviewer will comment on the quality of the prior work.) To what extent does the proposed activity suggest and explore creative and original concepts? How well conceived and organized is the proposed activity? Is there sufficient access to resources?

2. What are the broader impacts of the proposed activity? How well does the activity advance discovery and understanding while promoting teaching, training, and learning? How well does the proposed activity broaden the participation of underrepresented groups (e.g., gender, ethnicity, disability, geographic, etc.)? To what extent will it enhance the infrastructure for research and education, such as facilities, instrumentation, networks, and partnerships? Will the results be disseminated broadly to enhance scientific and technological understanding? What may be the benefits of the proposed activity to society?

3. Integration of Research and Education One of the principal strategies in support of NSF's goals is to foster integration of research and education through the programs, projects, and activities it supports at academic and research institutions. These institutions provide abundant opportunities where individuals may concurrently assume responsibilities as researchers, educators, and students and where all can engage in joint efforts that infuse education with the excitement of discovery and enrich research through the diversity of learning perspectives.

4. Integrating Diversity into NSF Programs, Projects, and Activities Broadening opportunities and enabling the participation of all citizens – women and men, underrepresented minorities, and persons with disabilities – is essential to the health and vitality of science and engineering. NSF is committed to this principle of diversity and deems it central to the programs, projects, and activities it considers and supports.

There are several final review criteria specific to this solicitation:

1. Possession of the scientific expertise and resources needed for tool development.
2. Possession of the scientific expertise and resources needed for the creation and analysis of databases on organizations and individuals.
3. Cohesion of technology, tools and data within each ”testbed”.
4. Documented outreach and dissemination plan.
5. Evidence of applicability to a broad range of sciences.
6. Quality of coordination plan.
7. Demonstration of scalability to, for example, additional organizations or other large-scale databases.
8. Evidence of long-term sustainability and impact.

References Cited

[1] The DARPA high productivity computing systems program. http://www.highproductivity.org/.

[2] Garageband. http://www.apple.com/ilife/garageband/.

[3] Hackystat developer services website. http://www.hackystat.org/.

[4] High productivity team workshop, January, 2005. http://www.highproductivity.org/Meetings-events.htm.

[5] Victor Basili, Marv Zelkowitz, Dag Sjoberg, Philip Johnson, and Tony Cowling. Ownership of data, testbeds, and artifacts. Technical report, University of Maryland/Fraunhofer Institute USA Technical Report ???, 2005.

[6] Kemal Ebcioglu, Vijay Saraswat, and Vivek Sarkar. X10: An experimental language for high productivity programming of scalable systems. In *Second Workshop on Productivity and Performance in High-End Computing*, San Francisco, CA., 2005.

[7] Martha S. Feldman and Brian T. Pentland. *Handbook of Organizational Routines*, chapter Issues in empirical field studies of organizational routines. Edward Elgar, Cheltenham, 2005 (expected).

[8] Andrew Funk, Victor Basili, Lorin Hochstein, and Jeremy Kepner. Application of a development time productivity metric to parallel software development. In *Second International Workshop on Software Engineering for High Performance Computing System Applications*, St. Louis, MO., May 2005.

[9] John Gustafson. Purpose-based benchmarks. *International Journal of High Performance Computing Applications*, 18(4), November 2004.

[10] Philip M. Johnson, editor. *Second International Workshop on Software Engineering for High Performance Computing System Applications*. Association for Computing Machinery, May 2005. http://csdl.ics.hawaii.edu/se-hpcs/.

[11] Philip M. Johnson, Hongbing Kou, Michael G. Paulding, Qin Zhang, Aaron Kagawa, and Takuya Yamashita. Improving software development management through software project telemetry. *IEEE Software*, August 2005.

[12] Philip M. Johnson and Michael G. Paulding. Understanding HPC development through automated process and product measurement with Hackystat. In *Second Workshop on Productivity and Performance in High-End Computing*, San Francisco, CA., 2005.

[13] Jeremy Kepner, editor. *Special Issue of the International Journal of High Performance Computing Applications on HPC Productivity*, volume 18, November 2004.

[14] Robert Numrich, Lorin Hochstein, and Victor Basili. A metric space for productivity measurement in software development. In *Second International Workshop on Software Engineering for High Performance Computing System Applications*, St. Louis, MO., May 2005.

[15] Nicholas Nystrom, John Urbanic, and Christina Savinell. Understanding productivity through non-intrusive instrumentation and statistical learning. In *Second Workshop on Productivity and Performance in High-End Computing*, 2005.

[16] Brian T. Pentland. *Variations in Organization Science: Essays in Honor of Donald T. Campbell*, chapter Organizations As Networks Of Action. Thousand Oaks, CA: Sage, 2005 (expected).

[17] Douglass Post, Kendall Richard, and Whitney Earl. Case study of the Falcon code project. In *Second International Workshop on Software Engineering for High Performance Computing System Applications*, St. Louis, MO., May 2005.

[18] Ram Rajamony, editor. *Second Workshop on Productivity and Performance in High-End Computing*. IEEE Computer Society, San Francisco, CA., February 2005. http://www.research.ibm.com/arl/pphec/.

[19] Burton Smith, David Mizell, John Gilbert, and Viral Shah. Towards a timed markov process model of software development. In *Second International Workshop on Software Engineering for High Performance Computing System Applications*, St. Louis, MO., May 2005.

[20] Larry Votta, Susan Squires, and Walter Tichy. What do programmers of parallel machines need? a survey. In *Second Workshop on Productivity and Performance in High-End Computing*, San Francisco, CA., 2005.