

# Research Proposal: Barriers to new user and new domain adoption of the XSEDE Cyberinfrastructure

Richard Knepper

School of Informatics and Computing / Pervasive Technology Institute  
Indiana University  
2709 E. 10th St, Bloomington, Indiana, USA  
Email: rich@iu.edu

**Abstract**—This research proposal proposes the examination of user attitudes about the Extreme Science and Engineering Discovery Environment (XSEDE). The XSEDE project supports basic research with a common system for making use of national cyberinfrastructure. The systems and infrastructure that make XSEDE useful for researchers are part of an actor network: these systems are socially constructed and they play their own part in the work of XSEDE, and in turn have an effect on the progress of basic research. I have completed previous work on the user relationships in the predecessor to XSEDE, the TeraGrid, and currently carry out participant observation with the management groups of the XSEDE project. By understanding the barriers to adoption of XSEDE by new researchers and new scientific domains, I hope to explore the linkage between resources (in this case computational resources) and scientific outputs.

## I. INTRODUCTION

**S**TARTING in 2001, the United States National Science Foundation (NSF) funded a distributed, high-performance computing project known as the TeraGrid, with the goal of supporting scientific research at the frontiers of current research. With over \$430M invested in the TeraGrid grants to date and more projects open for solicitation, the TeraGrid has created an ecosystem of interacting researchers, technologists, and administrators [1]. The TeraGrid was superseded in 2011 by XSEDE, which continues the implementation of the architecture started in the TeraGrid project with additional work towards making these large scale systems more flexible, usable, and open to a broader range of research agendas [2]. XSEDE has been awarded over \$64M to date for its management structure, with sites deploying resources based on NSF awards of as much as \$77M for the National Institute for Computational Science’s Kraken system, and \$55M for the Texas Advanced Computing Center’s Stampede system, to name the largest investments. The creation of national cyberinfrastructure in support of basic research is a resource-intensive effort: not only do the initiatives involved in XSEDE require substantial material investment, the centers that support these systems employ large numbers of staff to manage the systems, provide user support, adapt code to ever-larger systems and broader architectures, and reach out to the scientific community in order to bring in new researchers.

### A. XSEDE and outreach to new domains and users

The NSF mandates that the XSEDE project reach out to new disciplines and new institutions, including Minority Serving Institutions and Historically Black Colleges and Universities. Despite efforts to broaden the number of disciplines and institutions served in XSEDE, surveys of researchers indicate that only 30% of respondents feel that there are sufficient cyberinfrastructure resources available to meet their research needs. XSEDE’s quarterly resource allocations process is over-subscribed by roughly 3 times. In order for XSEDE to meet NSF’s recommendations to broaden its user-base, XSEDE needs to change the strategies used to recruit and develop users. Despite this charge from the NSF, non-traditional users of XSEDE still experience difficulties adapting and transitioning to the XSEDE project.

It is critical in the estimation of the NSF to provide a broad-use cyberinfrastructure framework for the support of basic science, and this has been emphasized in reports from the XSEDE project[3], as well as prominent researchers such as Richard Tapia and others. XSEDE includes in its management structure units for reaching out to new communities and has an emphasis on highlighting novel uses of XSEDE, but the current user base and user profile remains quite narrow. New systems on XSEDE advertise “Computing for the 99%”, but this paradigm has yet to be fleshed out, not to mention actual usage information.

### B. Motivation and research goals

This investment in infrastructure for research represents considerable outlay for the government in real terms. While not large compared to other government expenditures, these services provide the long-term innovative capacity for the country, and they influence the nation’s competitiveness and future technological development. XSEDE represents a system in which researchers have access to considerable resources, but only by adopting technologies as provided, and with access to resources governed by a peer-reviewed allocations process. In some ways this is like other resource-distribution mechanisms in science such as publications and grants, but the technological requirements to use and specificity of available

analyses on XSEDE resources means that different groups are more able to take advantage of resources.

The basic research questions in this study surround the impact that resources make on scientific work. XSEDE is a project intended to provide computational resources to scientists, no matter what their local resources may be. This research project is intended to reveal what these barriers to utilization of XSEDE resources are from the users' point of view, and to examine what these potential users of XSEDE say about their experiences, but also to analyze and understand the training and documentation materials provided for potential users, as well as observe some of the activities and NSF strategies for user recruitment. These findings will be examined along with the NSF's initiatives to generate user participation, with the goal of creating requirements for the next phase of the XSEDE project that can be incorporated into XSEDE's architectural processes: use cases and quality attributes. The products of architectural processes will include programmatic interfaces, training and documentation materials, and broader educational materials.

By understanding how researchers use resources, and how they compete for them, I intend to get more information about appropriate cyberinfrastructure and allocations methods to best serve the needs of the nation for basic research. By better understanding the users of XSEDE and their needs and relationships within XSEDE, successor infrastructures to this project can be built in order to do a better job supporting scientific research. My own previous research has focused on the management of XSEDE and relationships between supercomputing centers, and to understand the functioning of the project it is imperative to also see the project from the perspective of its client-constituents. By understanding linkages between resources and scientific outputs, it is possible to understand what kind of information is needed to drive science policy decisions as governments attempt to manage limited resources and still remain competitive.

## II. LITERATURE ON RESOURCES AND PERFORMANCE

In order to explore the connection between resources available to scientists and their productivity, it is important to understand the more general background of resources for scientific activities. Most commonly these are competitively awarded grants that provide scientists with the financial means to accomplish research goals.

While it may be argued that funding for cyberinfrastructure for research is qualitatively different from research practices, I draw upon the stream of research that examines performance management for research, noting a few major correspondences between research infrastructure and research itself. First and foremost, the large scale cyberinfrastructure that characterizes the TeraGrid and XSEDE projects is in itself treated as a computer science research project. Creating software that makes use of large scale systems efficiently and effectively, providing services that allow users to make use of cyberinfrastructure with ease, and creating links between large scale systems that allow them to be used in concert with each other

easily constitutes as significant challenge for computer science researchers. Secondly, the practice of providing research cyberinfrastructure for research is largely modeled after the process for obtaining grants for research itself. Proposals are created with broader impacts and scientific merits in mind, subjected to peer review, and projects are evaluated based upon the publications and reports they generate, as well services provided. Finally, there is limited evidence to suggest that infrastructure in support of research behaves and can be treated similarly to other infrastructure provided by other government agencies, thus our understanding of research funding is the closest analogue to assist in our understanding of dynamics driving cyberinfrastructure funding.

Performance management for grants is perhaps the most important difference between grant funded work and contracting work, and monitoring of scientific progress has singular difficulties. Partha and David [4] catalog the difficulties of economic evaluation of research: economic returns may come quickly or may take decades to realize, rights to intellectual properties are difficult to extract economic rents from (in fact restricting access to research may hamper further returns on initial investments), fundamental research progress may have dramatic and far-reaching impacts that are difficult to capture, and it is especially hard to forecast the success of any one particular research project. Hanson's [5] appraisal of focusing on effort as opposed to results has marked the transition from measuring research outputs towards measuring research processes. The standard operating procedure for evaluating both inputs (proposals) and outputs (scientific work) of grant-funded projects remains the peer review process. Garcia and Sanz-Menendez [6] discuss the context of peer review as metric of scientific research quite fully in their evaluation of competition in research initiatives. The authors begin tracing the path of peer review with the assertion that individual reputation and credit are central to the creation of the social structure of science, and that recognition by ones' peers is the foundation of legitimacy and leadership in a given field. Garcia and Sanz-Menendez note that the measurement of scientific production has long been based in volume and quality of scientific publications, but that these metrics cannot be separated from peer review. Peer review, despite some of its flaws outlined below, is not only the basic mechanism for ensuring quality of research, but also a critical factor in monitoring the efficiency of government investment in science. Peer review provides legitimacy to governmental bodies, and scientific work which has passed peer review has greater esteem in its scientific surroundings [6]. However, with the advent of new initiatives in government for assessing and monitoring of performance, peer review has had mixed fortunes as an evaluative tool for officials in charge of awarding research grants.

Shapira and Kuhlman [7] describe the growth in requirements evaluation of research projects as governments attempt to control costs and derive better benefits from programs, noting that there are significant issues to measuring performance in this area. Impacts of these programs tend to be diffuse, as do

costs, leading to difficulties in capturing all of the inputs and outputs. As research programs grow in complexity, including more disciplines and addressing broader problems, the evaluation of these programs must similarly become more complex. Government demands for continuous monitoring and program learning initiatives for research have led toward inclusion of subsidiarity, socioeconomic effects, and broader impacts into research evaluations. The increasing frequency of public-private partnerships for research also increases the complexity of program evaluation [7]. Partha and David [4] describe the the new attitudes towards measurement of research projects as a new economics of science, in which the previous free-market scientific workplace, characterized by scientists competing in the peer review process in order to gain funds and recognition is supplanted by a more interventionist government hand that is in the process of turning science toward more applied tasks. Government demands for better program evaluation both in the US and in Europe, as well as budgetary constraints from the recent economic crisis, have resulted in a call for more scrupulous examination of research performance.

The response to this call for increased evaluation and measurement of performance, has been variable at best. Cozzens [8], providing the context of evaluation in US research funding, describes the clash between the traditional evaluation tools for research, peer review and the journal selection process, and the new requirements based on the Government Performance and Results Act (GPRA) and increased requirements for management performance from the OMB. Peer review as the status quo for evaluation of science works in what Cozzens describes as an “autonomy-for-prosperity” model. Agencies support research activities in order to solve specific problems in an indeterminate amount of time, with limited oversight from Congress or agencies. Emphasis in evaluation is placed on the input end of the process, based on the quality and relevance of research proposals, and most importantly the accountability of this evaluation is placed on the research community, who is responsible for fairly making decisions, rather than on the researchers themselves to produce results to the general public. Guston [9] notes that peer review makes up a substantial amount of the selection process for research: \$37.7 billion or 86% of the reported total funding for research is merit reviewed. Applied research agencies, in contrast, have review processes based in personnel evaluation and budgeting that determine quality, although Cozzens, Bozeman, and Brown [10] note that there is a shift towards the competitive model of peer review even for these agencies. Peer reviewed grants are a feature of new federal research funding programs in the Department of Agriculture, the Environmental Protection Agency, and the Advanced Technology Program [9]. Peer review for research projects can happen both prospectively in the proposal selection process as well as retrospectively in evaluation [11], and have also been used as inputs in evaluating information for drafting regulation, creating state policies, and in evaluating courtroom decisions [9].

The peer review process conflicts directly with GPRA requirements for monitoring outputs of research, which ex-

PLICITLY focus on planning and achieving strategic objectives, rather than a culture of fairness in evaluation of proposals. As Cozzens [8] states, this “clashes with the traditional notion that the benefits that flow from research cannot be predicted in timing or content, but rather are visible only retrospectively”. Response to the new evaluation requirements has been met by providing measures that are generic and qualitative: outcomes such as “advances in knowledge” or “ideas, peoples, and tools”, or the NSF’s frequently sought-after “science nugget”, used to provide Congress with information about program success in a brief, easily-digestible package. As a result, such weak measures of evaluation lead to reinforcement of existing political forces, especially when coupled with another popular new metric of stakeholder input, which gives greater voice to those parties already engaged in the selection process [8]. Another approach to evaluation is to provide broad indicators of research progress: publications, funded research, and patents. Campbell [12] directly contrasts the peer review and indicator approaches finding that peer review results in complex but subjective evaluations of research work, while indicators are objective and easily quantified, but tend to be superficial in nature. Hagstrom [13] notes that peer reviewers frequently are able to identify the authors they are reviewing, or at least make educated guesses based on prior research and citation patterns. There is some evidence that researchers understand the peer review process and anticipate elements of it when drafting proposals. Knorr-Cetina [14] found in comparing proposals submitted for peer review to those without peer review that the style of the proposals changed rather than the content of the science inside. Furthermore, peer review is frequently conducted by established researchers, which leads to a problem in the assessment of new and innovative research directions, and relationships of mutual dependency that create self-reinforcing factions within scientific communities [12]. The world of the peer-reviewed scientist may be viewed as one mired in competition with other researchers first to get research proposals approved in order to get funding, and then to get the results of that research published.

Competition is in many ways the coordinating feature of scientific progress just as it is in economic activities. Hagstrom [13] describes the competition that takes place between scientists as specifically occurring when a scientist finds that her research in a particular area, on a problem not previously published, has been beaten to publication by another researcher. This form of competition may be extended to include being passed over in favor of another researcher in the grant selection process. Latour and Woolgar [15] established research funding as a vital part of researchers’ credibility and reputation with other scientists. Garcia and Sanz-Menendez [6] sum the idea of competition up well: “Thus, competition for funds is an essential mechanism in the cognitive functioning of research, articulated in the credibility cycle, and a vehicle for relationships between science and government”. Competition between researchers has a number of valuable features that aid scientific development. Competitive publication practices mean that additional researchers may be working on the

same problems, which ensures an abundant supply of possible investigatory techniques and results. Competition drives hard work on the part of the competitors to outdo each other. Finally, competition reduces the risk of dilatory publication, and it encourages differentiation and innovation as scientists attempt to identify new problems to explore [13]. While competition should promote the best quality research, issues have been identified with competitive processes for publication and funding that may slow the progress of science. Competition thus has a complex relationship with peer review. Laudel [16] notes that the competitive process may have impacts to the course of science as scientist averse to risk select other research topics in order to avoid competition and increase favor in peer review, promote mainstream or existing research techniques in order to be more competitive with particular review boards. Reports from leaders in grant-funded research centers find that competitive resubmissions for funding has a disruptive effect on getting the work of the center done [17].

With the understanding that research is driven by competitive processes, as scientists attempt to establish a track record that allows them to build capital to drive further research, and reputation that allows them to secure that research, a number of questions arise prompting further investigation into the workings of grant funded science. Firstly, what role does scale play in the competitive process? A number of authors have investigated the shift from individually-centered projects to larger labs and research centers. If scientists are engaged in competition for resources and prestige on the individual level, what forces do these interactions exert on the organizations that they work within? Does individual competition influence the structure and performance of these organizations? Secondly, what is the interplay between competition and collaboration in scientific projects? Federal agencies in the US and abroad have invested in funding projects to take on "Grand Challenge" projects such as the Large Hadron Collider, the caBIG project for cancer research, and the Laser Interferometer Gravitational-Wave Observatory (LIGO), which incorporate researchers from multiple institutions and are multidisciplinary investigations of research questions. Collaboration is an essential element of these large-scale scientific projects. However, Edwards notes that collaborative work is frequently hindered by "friction" of various types, categorizing "data friction" as issues that keep researchers from being able to easily exchange and manipulate data and "computational" friction which keeps scientists from easily making use of supercomputers or transferring work between different supercomputing sites [18], and later positing a "collaboration friction" which is the effort required for scientists in multiple disciplines and or differing backgrounds to work together in order to achieve collaborative success [19].

Finally, how can government agencies structure their programs and investments to take advantage of these factors? The NSF and to some lesser extent, the NIH, are under considerable pressure to reduce funding outlays and to provide rigorous performance management indicators for the scientific research done under their auspices. Are there features of

collaborative and competitive research that allow the grantors to get better results and better science faster?

### III. PREVIOUS RESEARCH EFFORTS

As a student of social informatics and of public management, with an interest in science policy, I have followed the development of the Supercomputing Centers program through the TeraGrid project and the XSEDE project. As an administrative manager in Indiana University's Pervasive Technology Institute, my duties include a management role within the XSEDE project, and this combination of theoretical, empirical research and access to the internal workings of the organization represent an opportunity for participatory research that is an extremely compelling case. In my efforts to get a better understanding of the TeraGrid user community, I conducted a social network analysis of the TeraGrid users and project allocations [20], noting the prevalence of traditional "hard science" disciplines, but also finding evidence of collaborations that span multiple fields of science. Since the end of the TeraGrid project and the advent of the XSEDE project, I have engaged in a case study [21], which characterizes the XSEDE project and similar advanced cyberinfrastructures as "living infrastructure" with qualitatively different attributes to traditional scientific instruments. These cyberinfrastructures have their own intents and goals that may or may not be aligned with their users' and the relationship between large cyberinfrastructure organizations and users becomes a collaborative arrangement in its own right.

Currently I am also working on a project with Katy Börner in the Department of Library and Information Sciences at Indiana University to characterize resource utilization on the XSEDE project as it relates to publication, in an effort to understand a general utilization-to-publication ratio, but also to examine whether different supercomputing centers and systems represent different research output profiles as well as the possibility of different fields of science having different resource/output functions.

### IV. HYPOTHESES

I hypothesize that there will be a number of barriers that are already well-described in documents such as the NSF's Advisory Committee on Cyberinfrastructure's Campus Bridging Task Force report, such as: difficulties in moving large amounts of data from its place of collection to computational systems for analysis difficulties in getting analyses to run on different resources unfamiliarity with national cyberinfrastructure frameworks and utilities I also hypothesize that in discussion with users, it will become evident that there is a lack of fit between the resources available via XSEDE and the new domains of interest. The XSEDE project and its predecessor, the TeraGrid, are rooted in the highly computational "hard science" disciplines such as high-energy physics, astronomy, and molecular dynamics. The new domains of interest to the NSF are slowly building computational emphases, but these techniques are often reliant on significantly large amounts of data, and analysis tools are developed on personal

computers, rather than the codes developed in the era of centralized computing facilities. These analyses include "Big Data" techniques—computational social sciences, for example—but also a significant number of life science analyses.

## V. METHODS OF ANALYSIS

A number of methods appear to make sense for investigating user benefits and difficulties taking advantage of XSEDE. XSEDE management already commits significant effort to improving practices within the project, from an in-depth architectural process to external evaluators who conduct general surveys of users and employees, as well as interviews in order to get more in-depth understanding, and case studies of individual initiatives within XSEDE. Leveraging the work of the external evaluation team can provide a considerable amount of information and indicate areas that require further inquiry.

### A. Document Analysis

In order to understand the current underlying body of knowledge, I propose to examine existing reports carried out by the external evaluation team, which has produced from 12-20 reports each project year [3]. This document analysis will be focused on user-centered evaluations (items such as the staff climate report will be excluded). Specific areas of interest will be the identification of mechanisms that support successful research activities as well as resource allocation schemes that have either facilitated projects or made them more difficult to carry out.

### B. Survey and Interview Activities

With some initial understanding of the existing data collected on XSEDE user activities, making use of the materials already present, I intend to conduct a survey of users with specific focus on sharing and utilization of resources, in order to understand how resources are obtained and allocated within projects. In order to understand the issues of new domains and new user populations—which may not easily conform to XSEDE's practices for obtaining access to and using resources, I will elicit proposals from the XSEDE allocations committee which were not approved, looking for proposals which were not turned down for lack of preparation or scientific merit, but for lack of fit to the project or for other issues. I will interview these researchers in order to understand how they planned to make use of XSEDE and what issues they felt kept their proposal from receiving an allocation of resources. These interviews will be conducted by teleconference or videoconference by one to one arrangement.

I also plan to conduct close interviews with members of the XSEDE "Campus Champion" program, who are local users who volunteer to assist others in making use of XSEDE by training and facilitating research activities. These individuals frequently help new users address problems getting started conducting analyses on XSEDE, and they often have some research component of their own to pursue.

### C. Participant Observation

In concert with the survey and interview activities described above, I intend to participate in training activities for new users, webinars and targeted outreach events that address user needs, and to engage in other activities, in hopes of observing users in action as they learn, examine, test, and adapt to the XSEDE environment. Events for participant observation of users will include the upcoming Linux Clusters Institute Workshop<sup>1</sup>, the XSEDE annual conference meeting, and online training and webinar events held by XSEDE, Ohio Supercomputing Center's HPC University, and Cornell University's Virtual Workshop program.

## VI. POTENTIAL BENEFITS OF RESEARCH

The current climate for science policy in the United States, but also in other countries is a problematic one. Support for basic research is being eroded: not long ago it was common to hear the phrase "flat is the new doubling" when speaking to research center leadership, this has changed to "minus five percent is the new flat". If the US and other nations are to ensure competitiveness and continued innovation, judicious use of resources and a strong understanding of the relationship between resources supporting research and scientific outputs is required. Scientific productivity is a problematic area to measure and there will be many different pieces of research that inform a larger stream, these will include studies of scientometrics and bibliometrics, but there is also significant work on large technical systems [22], [23] and the social element of technological systems [24], [25] that will contribute to our understanding of the social elements of scientific research.

This study will benefit our understanding of the linkage between access to resources and scientific outputs. In order for organizations such as the NSF to successfully support basic research, a better understanding of the effect of resources on research outcomes is required, especially in the case of shared resources such as the XSEDE project, which provides a long-lived scientific cyberinfrastructure service. XSEDE and its service provider units represent significant investment on the part of the National Science Foundation and making sure that the most effective distribution of resources can be made is critical to the longevity of the NSF and its mission. In a broader sense, understanding how centralized resources for research are best implemented in a general sense will provide better strategies for makers of science policy to present effective and equitable distribution of resources that ensures scientific progress and competitiveness in general.

## REFERENCES

- [1] NSF, "Nsf 07-28 cyberinfrastructure vision for 21st century discovery," 2007.
- [2] J. Towns, "Xsede: extreme science and engineering discovery environment," 2011.
- [3] N. S. F. OCI-1053575, "Xsede program years 1-3 comprehensive report," tech. rep., XSEDE, 2014.

<sup>1</sup><http://linuxclustersinstitute.org>

- [4] D. Partha and P. A. David, "Toward a new economics of science," *Research Policy*, vol. 23, no. 5, pp. 487 – 521, 1994. Special Issue in Honor of Nathan Rosenberg
- [5] R. Hanson, "Patterns of patronage: Why grants won over prizes in science," University of California, Berkeley, p. 11, 1998.
- [6] C. Garcia and L. Sanz-Menendez, "Competition for funding as an indicator of research competitiveness," *Scientometrics*, vol. 64, no. 3, pp. 271–300, 2005.
- [7] P. Shapira and S. Kuhlman, eds., *Learning from Science and Technology Policy Evaluation*. Cheltenham, UK: Edward Elgar, 2003.
- [8] S. Cozzens, *Frameworks for Evaluating S&T Policy in the United States*, ch. 4, pp. 54–64. Edward Elgar, 2003.
- [9] D. Guston, *The Expanding Role of Peer Review Processes in the United States*, ch. 6, pp. 81–97. Edward Elgar, 2003.
- [10] B. B.-B. E. Cozzens, S., "Measuring and ensuring excellence in government laboratories: Practices in the united states," tech. rep., *Canadian Council of Science and Technology Advisors*, 2001.
- [11] N. Science and T. C. (NSTC), "Assessing fundamental research," 1996.
- [12] D. F. Campbell, *The Evaluation of University Research in the United Kingdom and the Netherlands, Germany and Austria*, ch. 7, pp. 98–131. Edward Elgar, 2003.
- [13] W. O. Hagstrom, "Competition in science," *American Sociological Review*, vol. 39, no. 1, pp. pp. 1–18, 1974.
- [14] K. Knorr-Cetina, *The Manufacture of Knowledge: an Essay on the Constructivist and Contextual Nature of Science*. Oxford: Pergamon Press, 1981.
- [15] B. Latour and S. Woolgar, *Laboratory Life. The Construction of Scientific Facts*. London: Sage Publications, Ltd., 1979.
- [16] G. Laudel, "The art of getting funded: how scientists adapt to their funding conditions," *Science and Public Policy*, vol. 33, pp. 489–504, August 2006.
- [17] L. Smarr, "Hpcwire: The good, the bad and the ugly: Reflections on the nsf supercomputer center program.," 2010.
- [18] P. N. Edwards, *A vast machine: Computer models, climate data, and the politics of global warming*. MIT Press, 2010.
- [19] P. Edwards, M. S. Mayernik, A. Batcheller, G. Bowker, and C. Borgman, "Science friction: Data, metadata, and collaboration," *Social Studies of Science*, p. 0306312711413314, 2011.
- [20] R. Knepper, "The shape of the teragrid: analysis of teragrid users and projects as an affiliation network," in *Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery*, p. 54, ACM, 2011.
- [21] R. Knepper, "The xsede project: A living cyberinfrastructure." In Press, 2014.
- [22] T. P. Hughes, *Rescuing Prometheus: Four monumental projects that changed our world*. Vintage, 2011.
- [23] T. P. Hughes, *Networks of power: electrification in Western society, 1880-1930*. JHU Press, 1993.
- [24] W. Bijker, T. Hughes, and T. Pinch, eds., *The Social construction of technological systems: new directions in the sociology and history of technology*. MIT Press, 1987.
- [25] D. Mackenzie and J. Wajcman, eds., *The social shaping of Technology*. Open University Press, 1999.