Three Critical Matters in Big Data Projects for e-Science

Different User Groups, the Mutually Constitutive Perspective, and Virtual Organizational Capacity

Kerk F. Kee
Department of Communication Studies
Chapman University
Orange, California, USA
kerk.kee@gmail.com

Abstract—**This paper discusses three critical matters in big data projects for e-science, in which computer simulations, computational visualizations, and mathematical modeling are practiced as powerful methods to advance science. More specifically, three different user groups are identified, and their unique situations elaborated for developing project methodologies sensitive to their differences. Furthermore, 'technology development and technology use,' as well as 'technology and organizing' are two parallel processes argued as mutually constitutive within the individual pairs. Finally, project methodologies should include organizational capacity assessment and capacity building as strategic components to increase big data projects' likelihood of success in producing intended outcomes.**

Keywords-big data projects; e-science; cyberinfrastructure; technology implementation; innovation diffusion.

I. INTRODUCTION

Since the official establishment of the National Science Foundation's (NSF) Office of Cyberinfrastructure in 2006 [1], and its recent reinvention into the Division of Advanced Cyberinfrastructure in 2012 [2], there has been a significant investment and technological development in cyberinfrastructure (CI) to enable data-intensive and simulation-based e-science [3, 4]. The use of big data simulation for scientific investigation, which is embedded within the CI framework, has been hailed as "one of the most successful modern methods for experimental scientific discovery" [5, p. 30], an approach as equally powerful as theoretical and experimental research. Behind all the investments is the vision that effective CI diffusion (which includes the non-linear stages of design, development, adoption, use, implementation, and integration) will revolutionize science and engineering research and education in the U.S. [1, 6, 7, 8], keeping America a competitive leader in the global landscape of the STEM (science, technology, engineering, and mathematics) fields.

However, big data projects for e-science, in which computational tools and cyberinfrastructure are developed, face a unique set of challenges. These projects are usually inter-disciplinary, multi-institutional, funded on a limited basis (i.e., three to five years by NSF and other federal agencies), with the goal of advancing science (not directly the technologies), with a mix of full-time and part-time participants all working on a portion of their employment

efforts in creating experimental and specialized tools to harness big data for grand challenges in science.

Due to their inter-disciplinarity, creating a common language to facilitate understanding among diverse experts is difficult. Moreover, disciplinary diversity naturally leads to varying perspectives on project goals, which can make consensus and team cohesiveness hard to achieve. As the projects are multi-institutional, some participants are collocated but most are not. Therefore, an imbalanced mix of geographic proximity and distance challenges frequent and effective communication during project execution.

Furthermore, dispersed participants in a big data project do not depend on the project's principal investigator (PI) for performance evaluation and promotion within their respective institutions. However, there is a need for collaboration and contribution from everyone involved. This means directing a dispersed multi-institutional project involves a high transaction cost, in terms of managing time delay and exercising interpersonal influence, higher than that of a traditional brick-and-mortal corporation.

When the funding runs out, and the PI and co-PIs are not able to secure continuing funding, the technology development comes to a halt, and the participants disband and look for other partnerships and funding opportunities. The prototype technologies, if left online in an open source fashion, may be adopted and adapted by other scientists. But if they are overly niche and specialized, the usage beyond the initial purpose may be limited. Then the big data project becomes a one-off effort.

In addition, big data projects for e-science are made up of academics (i.e., professors, post-docs, scientific computing staff, graduate and undergraduate students at traditional universities), whose full-time and main responsibilities are not technology development and implementation for big data analyses. They do big data projects because they are inherently motivated to advance science. However, their attention to big data projects are distracted by a range of teaching responsibilities, service work, other projects' demands, and course work for some.

II. THE NEEDS TO DEVEOP NEW PROCESS METHODOLOGIES

Big data projects are springing up across various fields and industries. However, Saltz [9] points out that little is

known about the tool development methodologies, virtual team processes, and short-term organizational framework in order to facilitate, guide, and support big data projects towards effective, productive, and successful outcomes. In fact, Bird, Jones, and Kee [10, p. 41] maintain, "Researchers form VOs [virtual organizations] to collaborate, share resources, and access common datasets via the grid infrastructure," however, "[s]ometimes e-Science projects fail not because of the technologies but because of problems associated with organizing and managing grid infrastructures through VOs" (p. 41). The sociotechnical complexity captured by cyberinfrastructure as the platform, virtual organization as the process, and e-science as the enterprise presents a multifaceted case for studying big data projects. In other words, cyberinfrastructure, virtual organizations, and e-science are inextricably bound in big data projects, which need new and innovative methodologies to support them.

Saltz [9] argues that as big, heterogeneous, and multi-scale data continues to grow at an exponential rate, there is an increasing need to understand how to study and develop project methodologies to support these big data projects. Therefore, I argue that it is critical to develop a deep understanding of three topics often overlooked by organizational researchers of big data projects and e-science. In this paper, I focus the discussion on the context of e-science. In other words, the critical matters are most salient during the design, development, adoption, use, implementation, integration, and the ultimate diffusion of computational tools for big data analytics, simulations, visualizations, and mathematical modeling in science and engineering research.

First, there is a need to categorize users of computational tools in big data projects. By distinguishing the different user groups, we can better identify their needs and unique situations for designing project methodologies. Second, there is a need to better theorize the relationship between 'development and use,' as well as the relationship between 'technology and organizing' in big data projects for e-science. A more sophisticated and accurate perspective would help identify strategies to develop innovative team methodology and improve the virtual organizations for these projects. Third, there is a need to focus on defining organizational capacity and develop capacity building strategies for the implementation and adoption of computational tools in big data projects. While many participants in these projects are highly motivated, the lack of organizational capacity in various areas presents serious threats to projects' success. I will elaborate on the three critical topics in the rest of the paper.

III. Categorizing Three Groups of Technology Users

Lee, Bietz, and Thayer [11] argue that the traditional conceptualizations of technology 'developers' and 'users' do not fully capture the participants' roles in e-science

projects and stakeholders of cyberinfrastructure systems. In response to this argument, Kee and Browning [12, 13] identify three groups of users of computational tools in big data projects for e-science: scientist-developers, co-producers, and pure users of existing tools. First, *scientist-developers* are domain scientists who have mastered knowledge and skills in computational science. They (along with their post-docs, graduate and undergraduate students) are able to develop prototype computational tools (with varying qualities) to investigate their own research questions in e-science. In this case, the developers of computational tools and the users of early prototypes are the same people in the big data projects. They fully understand the research questions of the science and how big data will be used for the analyses, while keeping in mind the possibilities and limitations of the computational tools they develop. However, due to the practical challenges in time and background knowledge for mastering expertise in both domain science and computational science, individuals who are full-fledged scientist-developers are rare to find in traditional academia. Most scientist-developers have varying degrees of domain and computational expertise.

Second, *co-producers* are domain scientists who collaborate with computational technologists on joint projects that actively develop in-house and use custom-made computational tools in big data projects. In fact, the label of 'co-producers' applies to both the scientists as users and technologists as developers because they have to collaborate closely in the same projects. The computational tools developed have no immediate users, except for the domain scientists on the projects. The tools are often developed to serve the niche computational needs of the domain scientists involved. Without the scientific research questions of these close knit users, the big data projects would probably not have been funded by the NSF. This is because NSF is charged to fund science (and today doing science requires the appropriate technology), and often not the technology for technology's sakes [4]. Although the label of 'co-producers' appears to suggest equal participation from both the domain scientists and computational technologists, often the co-production of computational tools is ultimately for the purpose of advancing the domain science funded in the big data projects.

Third, *pure users of existing tools* are domain scientists who simply adopt and use computational tools and prototypes that either scientist-developers or co-producers share with the scientific community in an open source fashion. Many of the pure users do not know how to develop a tool, and they do not have collaborators in computational sciences to help them develop an in-house custom tool. Majority of the computational tools developed by the first two groups of users (i.e., scientist-developers and co-producers) are open source, partly because the open-source philosophy sits well with the academic and scientific computing community. Tools are open source also because

NSF funding is tax payer's money, thus, not appropriate to help grantees develop proprietary and commercial tools for personal financial gains. Furthermore, due to the short-term nature of big data projects for e-science (i.e., often funded for three to five years), the open source approach allows the experimental tools to be adopted and further developed by future new users, if these users find the prototypes useful and/or fitting for their own e-science and big data projects.

However, these pure users come to the tools later, and their adoption implies adaptation to the way the tools were initially developed, and how subsequent scientists and technologists shaped the tools as an open source community. Their interaction with the tools is primarily technology integration, but they do not always have prior experience and tacit knowledge to successfully adopt and integrate the tools. Moreover, they also have to take into consideration the long-term improvement, sustainability, and future diffusion of the tools before they heavily invest in integrating the tools into their big data projects, as they will depend on other users and developers to keep the tools active and updated for use over time.

These three groups of users and stakeholders work in cyberinfrastructure supported virtual organizations, but they interface within groups and across groups through a range of interpersonal and mass communication in person and via communication technologies. Sometimes feedback from the third group of pure users cycles back to the scientist-developers, contributing to the next iteration and revision of the adopted tool. Some other time, a group of co-producers (scientists and technologists) in a single VO may share a new computational tool online with the scientific community as an open-source technology, with the hope that the tool will get picked up by other scientists within and across scientific domains. When a scientist-developer leaves his/her own tool behind and adopts this new tool, it may be modified and improved by the scientist-developer and his/her group. Therefore, in both examples, development continues through the ongoing diffusion process.

An important aspect of the development and use of computational tools is that they depend on the successful formation and operation of effective virtual organizations and big data projects that pull together dispersed scientists, technologists, computational tools, big data, remote instruments, and high-performance computing resources to investigate new scientific questions that are not possible to investigate in local and self-contained laboratories using individual datasets and regular commercial computing systems [14]. Due to the dynamic nature of the enabling technologies described, these virtual organizations and big data projects are also dynamic. In this paper, the technologies and the virtual organizations are conceptualized as mutually constitutive of each other; the technologies and the virtual organizations cannot exist without the other, and it is their co-occurrence that sustains their mutual existence in big data projects. Therefore, the case of virtual organizations in e-science presents an opportunity to consider the mutually constitutive perspective to guide project methodologies.

IV. THE MUTUALLY CONSTITUTIVE PERSPECTIVE

Leonardi [15] proposes a theoretically sophisticated and empirically complex approach to study technology and organizations as mutually constitutive. He argues that past literature has inaccurately drawn a demarcation between 'technological development and technological use' at one level, and 'technology and organizing' on another. These two observations capture the essence of virtual organizations surrounding big data projects for e-science. These two false demarcations will be elaborated below.

A. Development and Use

First, Leonardi argues that the separation of technological development (i.e., object) and technology use (i.e., practice) separated by the act of implementation is empirically misleading and theoretically disadvantageous. The assumption that the development of a technology has come to an end when its use in the organizational setting begins at the point of implementation is a false assumption. He states, "Mounting evidence to the contrary suggests that treating implementation as a point at which the development of a technology stops and its use begins is empirically inaccurate" (p. 279). In fact, big data projects in science are uniquely suited for studying how the development of computational tools and their use take place simultaneously.

This first argument is true among all three categories of users identified by Kee and Browning [12, 13]. For independent scientist-developers, they simultaneously develop their own computational tools while putting the tools into use for their science. In this case, the developers and the users are the same people. In the case of close knit co-producers, both domain scientists as users and computational technologists as developers remain in constant communication about the design and implementation of the computational tools in the same funded big data projects. If they want to get renewing funding for their projects, they need to demonstrate that their computational tools are successfully developed to serve the funded science. In these two cases, it is apparent that technology development and technology use are tightly bound, recursive, and occurring as simultaneous.

A slightly extended feedback cycle may be experienced in the case of the third group, who are pure users of existing computational tools. However, when they experience difficulty, they often contact the developers and/or post their questions in an online open forum to seek help from the community of scientists as these tools are often open-source technologies. Feedback and solutions will directly impact the development and use of these existing tools. As soon as updates and changes are released, they are being implemented immediately by the larger community of users. Therefore, the blurring boundary between technology development and technology use is again true in this last group of users.

In all three cases, the development and use of computational tools in big data projects for e-science are

mutually constitutive. Therefore, for big data projects for e-science I argue alongside Leonardi's position that to conceptually draw an artificial line between development and use demarcated by what he calls the implementation line, is inaccurate. In big data projects, implementation therefore should be conceptualized to link the simultaneous development and use of computational tools in practice.

This argument has implications for big data project methodologies in the sense that both users and developers must recognize the need to work closely in sprints, an approach similar to what Saltz [9] describes as agile software development methodology. The developers need to provide ongoing opportunities for users to test-use the prototypes. Then the users need to put the tools to use immediately, and provide the developers quick but useful feedback. At this point, the developer takes the feedback and creates the next version for testing. The iterative cycles should occur promptly, and the iterative cycles should be recognized and accepted as an inherent nature of big data projects, not something to be resisted by project members.

Furthermore, developers should hold the mindset to have a bias for actions – to give users an early prototype to test and critique, so the prototype can be further adapted to the actual needs and preferences of the immediate users. Sharing a prototype-in-progress is better than perfecting the tool well into the development, because if the tool does not fully fit the immediate users' needs and preferences, taking a step back later will be impractical due to timeline and/or budget. Then the tool will likely continue down a risky path that further deviates from the immediate users simply for reasons that could have been avoided by getting early feedback.

B. Technology and Organizing

Second, Leonardi argues that researchers have conceptually separated technology and organizing by mistake. When we study technologies in organizations or organizations supported by technologies, both constitute the same set of objects, people, and interactions. For example, when we look at a technology in an organization, the units of analysis include the object, and the people working with the object, and the people working with each other around or via the object. Similarly, when we look at the organization and the process of organizing enabled by technologies, the units of analysis include the people interacting with each other around the technology, interacting through the object, and the object itself. Therefore, by studying the objects, people, and interactions, we can fully understand the relationship between technology and organizing. This perspective is especially suitable for studying virtual organizations in e-science as sociotechnical systems.

This second argument is again true in the case of big data projects for e-science, as these projects are often sustained by virtual organizations supported by cyberinfrastructure resources, along with its computational tools,

supercomputing networks, big data, remote instruments, communication technologies, human experts, etc. If we study a computational tool in these big data projects, the units of analysis include the tool, the scientist-developers, co-producers, and/or adopters, as well as their interactions with the tool and with each other around and via the tool. If we study the process of organizing enabled by cyberinfrastructure, we once again will analyze the people and their interactions with each other, via the tool in distributed collaboration, and the tool. Therefore, the mutually constitutive approach helps us recognize that the sociotechnical elements of technology and organizing in big data projects for e-science are entangled in a complex way.

By recognizing that the fundamental units of analysis are people, objects, and interactions, big data project methodologies can better focus on identifying the main people and objects, and designing strategic interactions to link the people and objects in productive ways. While the people, objects, and interactions are important for understanding cyberinfrastructure processes, organizational capacity is critical for understanding cyberinfrastructure diffusion.

V. ORGANIZATIONAL CAPACITY & CAPACITY BUILDING

Big data projects in science face a range of challenges due to the complexity and scope of CI as an innovation. Steward [16] defines CI as consisting of "computing systems, data storage systems, advanced instruments and data repositories, visualization environments, and people, all linked together by software and high performance networks to improve research productivity and enable breakthroughs not otherwise possible". In other words, CI is a complex system, involving a diverse network of interdependent technologies, remote instruments, big datasets, dispersed experts, diverse institutions, etc. [14]. Because CI is complex, involving a range of local and dispersed people, objects, and interactions, the methodology to support big data projects need to be sensitive to the sociotechnical complexity. The methodologies need to be different from approaching traditional projects that are brick-and-mortal with co-located people, objects, and face-to-face interactions. Understanding the complexity of CI is key.

A. Cyberinfrastructure as a Multi-Dimensional Innovation

Traditional innovation diffusion literature has tended to focus on studying a technology-based innovation as simply a material object for adoption and use. CI represents a new generation of innovations that is multi-dimensional, in that it involves the *material objects* (such as networks, hardware, software, big data, etc.), *organizational practices* (such as co-production of project-driven computational tools, distributed collaboration of multi-disciplinary experts, etc.), and *philosophical ideologies* (such as the belief that mathematical modeling, computer simulation, and computational visualization of big data is as valid a research approach as the traditional theoretical and experimental approaches, the belief that developing computational tools is

fundamental to the conduct of science with big data, etc.). Thus, the case of CI's multi-dimensionality presents an opportunity to create innovative project methodologies for big data.

B. Cyberinfrastructure as a Dynamic Innovation

Furthermore, CI's implementation and diffusion, especially the computational tools embedded in the larger platform, is not well understood. Traditional innovation diffusion literature has tended to focus on studying *static* innovations (i.e., pre-designed, mass-produced, bought off-the-shelf, and used-as-instructed after full product development, so the technologies remain static while the adopting organizations undergo new changes during implementation) within a social system. This traditional assumption can be seen as the result of the mistake Leonardi pointed out – technology development is assumed to have come to an end when technology use begins in organizations. However, computational tools represent a new generation of technologies that are *dynamic*, in that they are user-driven, custom-made, produced on demand, permanently beta, and more importantly, being put to use while simultaneously being developed in CI-enabled virtual organizations, which constantly co-evolve along with the technologies.

According to the paper by Saltz [9], based on studies that he reviewed, more than 50% of big data projects never get completed. He states, in fact, many more big data projects fail to achieve their project objectives. In e-science, big data project methodologies should be conceptualized not simply to consider how to 'run' a big data project, but to also assess and build organizational capacity in order to support the *multi-dimensional* (i.e., material objects, organizational practices, and philosophical ideologies) and *dynamic* (i.e., user-driven, custom-made, produced on demand, permanently beta, and being implemented in co-evolving virtual organizations) natures of CI.

While many participants in big data projects are highly motivated, the lack of organizational capacity in various areas present serious threats to projects' success. These areas include, but not limited to, stable funding, empowering policy, domain knowledge, computational expertise, technology training, high performance computing resources, collaboration technology skills, communication competence, personnel and human resources, critical support, etc. Furthermore, the organizational capacity necessary for big data projects are not geographically bound. Instead, they have to be appropriate for virtual organizations, making them unique as 'virtual organizational capacity' (VOC). Once an area (or more) of VOC is identified as lacking, strategic capacity building activities should be implemented in the big data project methodology to increase the VOC of the projects.

Linking the argument of VOC back to the three groups of users discussed earlier, the following observations can be made. First, scientist-developers require the capacity of

stable funding in order to sustain an appropriate period of technology development and use. Moreover, they need the capacity of computational expertise to enhance the big data projects. Both of these capacity factors are often compromised due to their dual roles, as they have to self-teach and then keep up with computational sciences in addition to keeping up with their own domain sciences. Second, co-producers (or co-producing users) need the capacity of communication competence to help them articulate their tool requirements to developers. The co-producing developers need the capacity of domain knowledge to understand the science being served by the tool they design. Users' communication competence and developers' domain knowledge are essential for facilitating coherent goals, mutual understanding, and collaborative culture in the projects.

Third, pure users of existing tools need the capacity of technology training and critical support to integrate a tool into their existing science, workflow, and infrastructure. For example, a computational tool developed by a group of scientist-developers in one VO may diffuse to another VO of pure users, as they simply adopt the existing tool. If these adopters do not possess sufficient capacity of technology training and critical support to implement the tool, the project will not likely succeed. Finally, all three groups need the capacity of empowering policy, high performance computing resources, personnel and human resources, and collaboration technology skills. Effective big data project methodologies needs to be sensitive to the range of factors that constitute VOC, so the methodologies can be designed to cater to the unique needs, challenges, and situations of each user group.

VI. CONCLUSION, DISCUSSION, AND IMPLICATIONS

In this paper, I advanced a few arguments about the sociotechnical complexity of big data projects for e-science. First, it is important to distinguish among the different groups of users in big data projects. More specifically, they are the scientist-developers, co-producers, and pure users of existing tools. Understanding their unique roles, backgrounds, prior experiences, needs, and situations would help in developing project methodologies appropriate to these differences.

Second, I presented the case of big data projects in which virtual organizations are formed to pull together a diverse range of people, objects, and interactions. More specifically, I explained the ways in which big data projects involve the simultaneous co-occurrence of technology development and use, as well as technology and virtual organizing. These two points suggest that project methodologies should design iterative interactions among a set of people and objects, likely via technological and digital connections.

Third, I defined cyberinfrastructure as a multi-dimensional and dynamic innovation, with which virtual organizations for e-science also co-evolve along with cyberinfrastructure. Project methodologies should involve

assessing the virtual organizational capacity of big data projects and implement strategic capacity building activities to increase projects' likelihood of producing the intended outcomes. More specifically, capacity assessment and capacity building should involve supporting the material objects, organizational practices, and philosophical ideologies necessary to design, implement, and integrate cyberinfrastructure and big data for e-science.

Given the discussion above, I propose that the following research questions to guide future explorations of big data projects for e-science:

- *How may agile practices (between technology development and use) be adapted to support big data projects?*
- *How can 'mutual constitutiveness' between technology and organizing in big data projects be empirically characterized?*
- *What specific types/factors of organizational capacity must be developed in order to enable successful operation of big data projects?*
- *How can specific types of organizational capacity be conceptualized and operationalized for organizational assessment?*
- *How can virtual organizations for big data projects be helped in understanding where they are lacking in capacity?*

Beyond the specific context of e-science, this paper also has implications for big data projects at the general level. First, the simultaneous and interactive development and use of technologies is not well understood, because most studies in the past have focused on static innovations. However, the trend is that new technologies today are often networked. This networkability allows developers to receive real-time error messages and constant feedback from users. Furthermore, this networkability also allows users to receive new updates from developers on a regular basis. This trend is beginning to make simultaneous development and use more common that in the past, making new technologies 'permanently beta' and open ended. College training of technologists (in computer and computational sciences programs) should prepare students to be comfortable with engaging in iterative cycles of development, instead of considering a technology development task complete when an assignment is turned in to the professor.

Moreover, big data project methodologies would be wise to include strategies for creating a common project understanding, facilitating constant communication cycles, and promoting synergistic collaborations. If one party does not see itself as an equal partner who needs to proactively engage the other group with trust and respect, the project is likely to struggle. As big data emerges as a complex and constantly evolving topic, the tools required for big data analyses will be user-driven, custom-made, produced on demand, permanently beta, and more importantly, being put

to use while simultaneously being developed in the respective virtual organizations, which constantly co-evolve along with the technologies. The organizational capacity necessary to support such dynamism is critical for all big data projects, not just those in e-science involving cyberinfrastructure.

Finally, as in the case of scientist-developers, data scientists may be required to play the dual role of domain experts (as users of big data for specific domain applications) and computational analysts (as programmers and experts in making sense of data at the computational level) simultaneously. Given the dual roles, the ability to garner sufficient organizational capacity across the levels of materials, behaviors, and ideologies present perhaps one of the most critical, if not the most critical, challenge(s). Projects can fail because 'domain expert-computational analysts' simply cannot keep up with the demand of maintaining two full time careers concurrently.

REFERENCES

[1] E. Seidel, J. Muñoz, S. Meacham, and C. A. Whitson, "A Vision for Cyberinfrastructure," Computer, vol. 42, Jan. 2009, p. 40.

[2] NSF. National Science Foundation Realignment Plans, 2012, http://nsf.gov/news/news_summ.jsp?cntn_id=125381.

[3] P. N. Edwards, S. J. Jackson, G. C. Bowker, and R. Williams, "Introduction: An Agenda for Infrastructure Studies," J of the Association for Information Systems, vol. 10, May. 2009, pp. 364-374.

[4] K. F. Kee, and L. D. Browning, "The Dialectical Tensions in the Funding Infrastructure of Cyberinfrastructure," Computer Supported Cooperative Work, vol. 19, Aug. 2010, pp. 283-308.

[5] V. Getov, "e-Science: The Added Value for Modern Discovery," Computer, vol. 41, Jan. 2008, pp. 30-31.

[6] D. E. Atkins, K. K. Droegemeier, S. I. Feldman, H. Garcia-Molina, M. L. Klein, D. G. Messerschmitt, et al., "Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure", National Science Foundation, 2003, http://www.communitytechnology.org/nsf_ci_report/.

[7] C. L. Borgman, H. Abelson, L. Dirks, R. Johnson, K. R. Koedinger, M. C. Linn, et al. "Fostering Learning in the Networked World: The Cyberlearning Opportunity and Challenge," Report of the NSF Task Force on Cyberlearning, 2008, http://www.nsf.gov/pubs/2008/nsf08204/nsf08204.pdf.

[8] NSF. "Cyberinfrastructure Vision for 21st Century Discovery," 2007, http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf.

[9] J. S. Saltz, "The Need for New Processes, Methodologies and Tools to Support Big Data Teams and Improve Big Data Project Effectivenss. IEEE Big Data Workshop on Methodoologies to Improve Data Projects (IEEE BIGDAYA 15), IEEE Press, in press.

[10] I. Bird, B. Jones, and K. F. Kee, "The Organization and Management of Grid Infrastructures," Computer, vol. 42, Jan. 2009, pp. 36-46.

[11] C. P. Lee, M. J. Bietz, and A. Thayer, "Research-Driven Stakeholders in Cyberinfrastructure Use and Development", the International Symposium on Collaborative Technologies and Systems (CTS), IEEE Press, May. 2010, pp. 163-172, doi: 10.1109/CTS.2010.5478514.

[12] K. F. Kee, and L. D. Browning, "Challenges of Scientist-Developers and Adopters of Existing Cyberinfrastructure Tools for Data-Intensive Collaboration, Computational Simulation, and Interdisciplinary Projects in Early e-Science in the U.S.", Data-Intensive Collaboration in Science and Engineering workshop, Computer Supported Cooperative Work (CSCW 12), Seattle, WA, 2012.

[13] K. F. Kee, and L. D. Browning, "Two Socio-Technical Gaps of Cyberinfrastructure Development and Implementation for Data-Intensive Collaboration and Computational Simulation in Early e-Science Projects in the U.S.," Mastering Data-Intensive Collaboration Through the Synergy of Human and Machine Reasoning workshop, Computer Supported Cooperative Work (CSCW 12), Seattle, WA, 2012.

[14] K. F. Kee, L. Cradduck, B. Blodgett, and R. Olwan, "Cyberinfrastructure Inside Out: Definitions and Influences Shaping Its Emergence, Development, and Implementation," D. Araya, Y. Breindl and T. Houghton, Eds, Nexus: New intersections in Internet research. New York: Peter Lang, 2011, pp. 157-189.

[15] P. M. Leonardi, "Crossing the implementation line: The mutual constitution of technology and organizing across development and use activities," Communication Theory, vol. 19, Aug. 2009, pp. 278-310.

[16] C. Stewart, "Indiana University Cyberinfrastructure Newsletter", 2007, http://racinfo.indiana.edu/newsletter/archives/2007-03.shtml.