Stakeholders in Cyberinfrastructure Development

Charlotte P. Lee, Matthew J. Bietz, & Alex Thayer
University of Washington, USA
[cplee,mbietz,huevos]@uw.edu

Abstract

Research has shown that failing to recognize and understand organizational subgroups, their cultures, and their reward systems can result in a failure of system adoption. Infrastructure building projects for science are complex forms of collaborative work that involve many subgroups. As part of an ongoing research project, we use ethnographic methods to explore the roles, categories, and relationships that are sometimes taken for granted in cyberinfrastructure research and development. We find a diversity of stakeholders and stakes in the development of a cyberinfrastructure for environmental genomics that transcends categories such as "users," "designers," and "community."

1. Introduction

Research on scientific collaboration, e-Science, and the development and use of large-scale cyberinfrastructures (CIs) is an expanding area of inquiry. CIs are distributed organizations supported by advanced technological infrastructures such as supercomputers and high-speed networks. These projects present special challenges to designers and developers because of their large-scale nature and long duration requirements for development and use, and because of their involvement at the "cutting edge" of science.

As science changes, CIs must adapt to new scientific questions and technologies, new user groups, and changing technical requirements. Human-centered design methods stress the importance of understanding the primary users' needs for a product or system, developing a set of requirements tied to those needs, and then designing to meet the requirements. CI projects, however, are comprised of diverse users with rapidly evolving needs. Additionally, the "users" are not always end users of the CI; policymakers, project funders, and other interested parties may contribute requirements or constraints.

Therefore, it is useful to revisit the relationship between users and developers in the context of the CI domain. A broader approach to the design effort is required, an approach that meets end-user requirements and aligns stakeholder interests.

2. Users, Stakeholders, and Community

The concepts of "user" and "stakeholder" are often defined in terms of their interaction with a particular information system [1, 2]. The precise role that stakeholders play in the design process depends in part on the prevailing design philosophy of a project team or organization. For example, the "designer as expert" perspective posits that designers "consider themselves to be the experts, and they see and refer to people as 'subjects,' 'users,' 'consumers,' etc." [3]. This approach often characterizes users as research subjects who undergo usability studies and provide interpretive data about design prototypes, and stakeholders as business decision makers who submit requirements to the designers as their only contribution to the design process. This distinction establishes an artificial boundary between the designer-expert and the users and stakeholders for whom he or she is designing.

In the current study, we approach users and designers as categories of stakeholders implicated in building CIs. We observe that:

- Categories like "user," "designer," etc., represent multiple, not unified, interests,
- Important interests transcend the boxes of "user," or "designer," and
- Organizations and projects may also be seen as stakeholders, a perspective that is aligned with Latour's understanding of translating interests in the production of science and with research in Participatory Design (PD).

The stakeholder perspective allows insight into the process of translating interests into infrastructure without assuming categories of actors that may or may not be salient in the moment of design. This approach also allows us to look more broadly at other categories of actors who may have a stake in the development of the infrastructure, and to investigate how those stakes are (or are not) represented in the technological artifacts and human structures of CIs.

3. Research Site & Method

CAMERA is a large-scale, multi-year project to provide CI tools, resources, and expertise for *marine metagenomics* research. Metagenomics is a "new science" that transcends a focus on individual organisms to study the genetic composition of populations of microbes [4]. The CAMERA project is one attempt to meet this field's computation and data requirements. CAMERA and its collaborators are developing specialized tools for data storage, analysis, and visualization. Anyone can gain access to most of CAMERA's tools and data at the project website.

We studied the development of CAMERA over a 2-year period. We have completed 19 in-depth semi-structured interviews with 13 members of the project. During 4 months of intense engagement, we observed 8 weekly group meetings of the development team, 6 scheduled subproject meetings, and numerous ad-hoc meetings. During this time, 1 of the authors also conducted more than 50 hours of unstructured on-site observations and had an assigned desk in the development team's work area.

We also turned our investigations to other potential stakeholders using a snowball sampling technique, resulting in an additional 14 interviews, attendance at 10 laboratory meetings of a metagenomics laboratory, and numerous informal conversations with individuals involved in metagenomics research.

Interviews were transcribed and, along with field notes and collected documents, analyzed in Atlas.ti software using a grounded theory approach [5].

4. Multiplicity of Stakeholders

Discussions of CI development have recognized that tensions and cultural differences between domain scientists and developers can be disruptive [6, 7]. With CAMERA, we find that this dichotomy is not sufficiently nuanced to explain the design process. Here, the "domain scientist" and "developer" categories are actually comprised of multiple sub-categories of stakeholders, all with different needs. In addition, certain stakeholders are involved in development but are neither domain scientists nor developers, while other project participants serve as both domain scientists and developers simultaneously.

In this section, we describe specific types of stakeholders and their interests, or "stakes." Our description is an illustration of the multiplicity and complexity of concerns that play a role in CI development, and is not intended as a comprehensive discussion of all possible stakes and stakeholders.

4.1. Blurring of Roles

Individuals who are associated with a CI often inhabit multiple roles that change as the situation demands. This dynamic holds true for domain scientists (users) and computer scientists (developers) on the CAMERA project, partly because the vision for CAMERA is that scientists will contribute original analysis tools and software along with the data they submit. As a result, the domain scientists who are "users" of CAMERA may also find themselves developing software that will be incorporated into the CAMERA system.

The CAMERA domain scientists recognize this fuzzy distinction between users and developers. The head of a microbiology lab told us:

> *I think at least the students and the post docs in the lab are hopefully getting some pretty good training in the computational sides of things. You know, everybody who comes in, the first thing that they're given is introduction to Unix and, you know, how to program in Perl books because it's going to be - you know, as necessary tools of the trade, at least in my lab, as, you know, the pipetters.(Alex)*

We also find an increasing number of "hybrid" individuals involved in CI development. For example, computational biologists are trained in both computer science and a particular domain. Similarly, bioinformaticists bring sophisticated statistical techniques to bear on biology questions. These interstitial categories are breaking down the user/developer barrier for CI development in incremental but profound ways.

4.2. Stakes Beyond Direct "User" Interaction

People who might typically be thought of as "users" often have a stake in the system that goes beyond their own direct interaction with the system. For example, one domain scientist told us that he supports the CAMERA project because it allows him to free up laboratory resources for other purposes.

> *My motivation is still about the same. We really need these tools outside of the labs and in a centralized system…. It's absurd for me to be running - I run a lot of computers and I hate computers a lot.(Christopher)*

For this scientist, CAMERA represents an opportunity for him to shift the burden of maintaining computers out of his lab.

Scientists might also gain indirect benefits from CAMERA's success. One scientist who is designing an analysis tool that may be included as part of the CAMERA system expressed a sense of his reputational stake in CAMERA:

It's going to be presented as a community tool…. I think these infrastructures like CAMERA are really important to organize all of the people and just being part of this project is good for me in terms of just networking with people and just learning more. And I can really see how if I put on my vitae later on… that I've been in touch, I have interacted with the CAMERA people, it would totally be a positive point.(Anthony)

Scientists who contribute data and tools to CAMERA find that their own reputation is bound to the overall success of the infrastructure.

Database systems like CAMERA also provide one route through which research community values and boundaries are established. Bietz and Lee [8] found that sequence databases act as "boundary negotiating artifacts." The design of the system can favor particular scientific questions and approaches. CAMERA will not only support individual scientists' own work, it may also help to raise the profile of metagenomic science. This could potentially lead to more funding, increased legitimization of the field, and greater reputation for early adopters of metagenomic techniques. Scientists have a stake in ensuring that their own approaches are supported by the system.

CAMERA's main source of funding is a grant from the Gordon and Betty Moore Foundation (GBMF), and like any funding agency they have a strong interest in the endeavor. Their concern extends beyond the success of the project itself. GBMF funds a number of "Moore Investigators," scientists who are given grants to pursue their own scientific projects. Collaboration with CAMERA is written into many of these grants, which often specify that the investigators must make their data available through the CAMERA database. GBMF is funding both the development and use of the infrastructure, giving it a complex but very important stake in the CAMERA project.

Other people and organizations may hold a stake in CAMERA even if they are not as directly involved in using or developing the CI. For example, CAMERA exists within a landscape of genetic databases, requiring at least some level of data sharing and compatibility with these other systems [8].

4.3. "Community" in CI Development

"Community" plays a significant role in how the project is defined by participants:

I think the overall goal for CAMERA is to really build a community around metagenomics, provide a data repository for metagenomic-specific data sets and the associated tools…. It's kind of a community for metagenomics data, people and resources. (Daniel)

CAMERA serves the metagenomics community at the same time that it is a tool to develop that community. CAMERA is both a community resource and a community itself.

The CAMERA development team relies on several techniques to understand the community. For example, the Scientific Advisory Board includes domain scientists who meet regularly and provide guidance for the project. The developers have also adopted a strategy of testing new features with a small number of individual scientists. These "early adopters" were seen not just as individual users, but as representatives of user communities.

Ribes and Finholt [9] looked at the concept of "community" within the development of another CI project, the WATERS Network. There, community becomes a category that is mobilized through conversation, community forums, surveys, and individual representatives. Community is used as "a short-hand for issues of representation, for example, ensuring inclusion and establishing a mandate—in other words, as a near synonym for what in the sphere of politics we would call a constituency" (p. 115).

Like Ribes & Finholt [9], we find that "community" takes on special significance in CI development. However, where they trace the "formation of a single community from heterogeneous beginnings" (p. 115), we find that, in the CAMERA project, "community" remains multiply-defined and malleable. At various times, participants employed "community" to refer to: all registered users, metagenomics researchers, researchers from another specific discipline (e.g. marine biologists), researchers who use genetic sequence data in their work, or even the broad "scientific community." Even when participants speak of "developing the community," there is an awareness that this does not necessarily signify a unitary construct.

So metagenomics is a very interdisciplinary field and that means including biologists and ecologists or chemists and computational biologists, bioinformaticians and then technical people, computational scientists, and also too developers, software developers and other engineers to develop other technology for; these are all a part of a community. (Emma)

CAMERA is explicitly part of a larger strategy to create new, transformative science. CAMERA participants use community "as a short-hand for issues of representation" [9], but the term remains usefully vague. Its fluidity becomes a resource for developers as they work to translate high-level project goals into technological features, supporting the production of technological artifacts within a fragmented and evolving design space.

At the same time, CAMERA's developers recognize their role in the ongoing development of the scientific community they want to serve:

> *I think the long-term vision of CAMERA is to both enable the creation of a new research community centered around the field of metagenomics, as well as be the key enabler at the center - the key computational and cyberinfrastructure enabler at the center of that community. (Michael)*

CAMERA is mandated to serve marine metagenomics, but this field will only be part of the eventual community of stakeholders. Other researchers who study other environments or use different methods are likely to use and contribute to CAMERA. Metagenomics itself will change, with not only new techniques and data sources, but also the involvement of new communities. The result will be an inevitable shift in the landscape of stakeholders.

5. Discussion

In CI development, it is useful to reframe "eliciting user requirements" as a process of discovering and understanding complex and continuously evolving sets of stakeholders and their interests. Focusing on stakeholders as we have outlined here helps to illuminate the diversity of CI projects, with their wide variety of purposes and organizational and social arrangements. The challenge for both the designer and CSCW researcher is to understand how stakeholders are arranged and how stakes are (or should be) prioritized in a particular context. Designing for stakeholders is but one small step toward recognizing and designing for the multiplicity of concerns and activities of CIs, and planning strategically for the continual emergence of stakes that must be continually aligned and realigned.

6. Acknowledgments

The authors wish to thank the various members of the CAMERA project, metagenomics researchers, and others who participated in this research. This work was supported by the National Science Foundation (IIS-0712994 & OCI-083860).

7. References

[1] D. Dinka and J. Lundberg, "Identity and role--A qualitative case study of cooperative scenario building," *International Journal of Human-Computer Studies,* vol. 64, pp. 1049-1060, 2006.

[2] J. A. Ross and S. Ben Jaafar, "Participatory needs assessment," *Canadian Journal of Program Evaluation*, vol. 21, pp. 131-154, 2006.

[3] L. Sanders, "An evolving map of design practice and design research," *interactions,* vol. 15, pp. 13-17, 2008.

[4] National Research Council (U.S.). Committee on Metagenomics: Challenges and Functional Applications, "New science of metagenomics: Revealing the secrets of our microbial planet," National Academies Press, Washington, D. C.2007.

[5] B. G. Glaser and A. L. Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New York: Aldine de Gruyter, 1967.

[6] S. L. Star and K. Ruhleder, "Steps toward an ecology of infrastructure: Design and access for large information spaces," *Information Systems Research*, vol. 7, pp. 111-134, 1996.

[7] B. F. Spencer, Jr., R. Butler, K. Ricker, D. Marcusiu, T. A. Finholt, I. Foster, C. Kesselman, and J. P. Birnholtz, "NEESgrid: Lessions learned for future cyberinfrastructure development," in *Scientific Collaboration on the Internet*, G. M. Olson, A. Zimmerman, and N. Bos, Eds. Cambridge, MA: MIT Press, 2008, pp. 331-347.

[8] M. J. Bietz and C. P. Lee, "Collaboration in metagenomics: Sequence databases and the organization of scientific work," in *ECSCW 2009: Proceedings of the 11th European Conference on Computer Supported Cooperative Work, 7-11 September 2009, Vienna, Austria*, E. Balka, L. Ciolfi, C. Simone, H. Tellioğlu, and I. Wagner, Eds. London: Springer-Verlag, 2009, pp. 243-262.

[9] D. Ribes and T. A. Finholt, "Representing community: knowing users in the face of changing constituencies," in *Proceedings of the ACM 2008 conference on Computer supported cooperative work* New York: ACM, 2008, pp. 107-116.