Research-Driven Stakeholders in Cyberinfrastructure Use and 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 investigate the difficulty of modeling stakeholders in the development of research-driven, large-scale scientific research and describe the importance of identifying stakeholders according to research questions in addition to organizations or workplaces.

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.

Our research site is a large CI project being developed to support marine metagenomics (also known as environmental genomics) research. The Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA) project, based at the University of California San Diego (UCSD), is accumulating an extensive database of metagenomics data and building tools to analyze those data. The CAMERA project is also actively involved in growing the nascent marine metagenomics scientific community. As part of an ongoing research project, we use ethnographic methods to describe and analyze the sources of design criteria in a CI development project. In particular, we find that CI developers must manage the needs of diverse stakeholders.

2. Science and Complex Collaborations

Failing to recognize and understand organizational subgroups, their cultures, and their reward systems can result in a failure of system adoption even in a relatively simple co-located organization [1]. Understanding the social landscape is even more important in CI-building projects, which typically involve multiple institutions, labs, investigators, funders, etc. The notion of human infrastructure helps us to understand organizational subgroups and their reward systems. The term "human infrastructure" was coined by the former director of the San Diego Supercomputer Center, Fran Berman [2], and has subsequently been used by Lee et. al [3] to explore the variety of forms that collaboration may take in the development of large-scale collaborations such as CI development.

Human infrastructure posits that complex infrastructures arise through complex interactions among networks, place-based organizations, groups,

and consortia. Human infrastructure does not conform to a particular collaborative structure (a team, organization, network, etc.) but can be comprised of multiple structures simultaneously. Structures can also change shape over time; these collaborative structures configure and are configured through the activities of infrastructuring. Clearly, collaborative infrastructure building for science presents a complex design space.

3. Users, Stakeholders, and Community

The concepts of "user" and "stakeholder" are often defined in terms of their interaction with a particular information system. The systems-based distinction made by Friedman, et al. [4], describes users as "direct stakeholders" who are "parties, individuals, or organizations who interact directly with the computer system or its output," while "indirect stakeholders" are "all other parties who are affected by the use of the system" (ibid). The role-based definition claims that users are defined by different aims and stakes connected to their professional role [5, 6]. Users are but one type of stakeholder.

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." [7]. 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. The participatory design approach is one of several such approaches that attempt to remove this boundary. Originally this approach focused on "users" and "designers," but increasingly the scope of the approach has expanded to include stakeholders as well [5].

In the literature on CI development, the popular notion of community has gained favor but still retains baggage from the user-designer dichotomy. In their work on the design of the Worm Community System, Star & Ruhleder [8] refer to the users and developers as being in separate camps that need to meet halfway. Others have noted that "domain scientists" and computer scientists have different motivations and cultures [9-11]. In CI parlance, "domain scientist" and "computer scientist" are used in much the same way as user and developer. The notion of community, however, has been used in more nuanced ways.

Ribes and Finholt [12] looked at the concept of "community" within the development of another CI project, the WATERS Network. They outline the ways that the WATERS participants worked to know and represent the scientific community's needs in the design process. Rather than just being a description of shared characteristics, 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).

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

- Each category of "user," "designer," or "community" represents multiple, not unified, interests,
- Important interests transcend the boxes of "user," "designer," or "community," 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.

4. Research Site & Method

CAMERA is a large-scale, multi-year project to provide CI tools, resources, and expertise for *metagenomics* research. Metagenomics is a "new science" that transcends a focus on individual organisms to study the genetic composition of populations of microbes [13]. Scientists using metagenomics techniques extract and analyze DNA taken from communities of microorganisms living in different environments. This is made possible by advances in high-throughput DNA sequencing along with new laboratory techniques that can extract DNA from the 99% of the world's microorganisms that

cannot be cultured in a laboratory. Access to new sources of data coupled with the increased speed and lower cost of sequencing have led to an explosion in the amount of available genetic data. Scientists need powerful computers and networks for data sharing and analysis in order to conduct metagenomics research and deliver novel biological solutions to important societal challenges in health care, energy, and the environment [14].

The CAMERA project is one attempt to meet this field's computation and data requirements, providing access to high performance computing clusters that currently run over 500 processors and provide more than 150 terabytes of data storage. CAMERA and its collaborators are developing specialized bioinformatics tools for data analysis. High-resolution, multi-monitor visualization walls are being deployed to metagenomics laboratories. Multiple computing and visualization sites are being connected through the high-speed OptIPuter network [15]. Anyone can register and gain access to most of CAMERA's tools and data at the project website (http://camera.calit2.net/). The CAMERA project is intended to serve as a model for other disciplinary sciences as they adopt CIs.

Metagenomics, however, is an emerging field, with committees debating over standards for metadata, developers creating new data analysis tools, funding agencies developing new grant categories, and scientists just beginning to identify themselves as metagenomicists. Furthermore, because of the sheer volume of the data, new techniques and technologies are being developed to generate, store, analyze, and disseminate data. This study presents a unique opportunity to examine the intertwined process of creating infrastructure and doing science.

CAMERA is a relatively new project, receiving funding since 2006. It affords study of the early years of a CI development effort. The participants in our study are all participating in CAMERA, although not all them are members of the development team. All participant names used in this paper are pseudonyms.

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. Six of the full-time developers have desks in this open-plan work area, and other team members have offices on the same floor. Observations included shadowing several of the team members, sitting in on casual conversations among the team, and noting general patterns of interaction among team members. Because much of the development work is highly technical and focused on the computer screen, sitting with the group provides the opportunity to ask developers about current work, or to provide explanations and context for events.

We also turned our investigations to other potential stakeholders using a snowball sampling technique. We asked development team members to list their collaborators and suggest any other individuals we should interview. We also attended metagenomics-related conferences and workshops, where we recruited study participants. These activities resulted 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, coded and analyzed in Atlas.ti software using a grounded theory approach [16].

5. Multiplicity of Stakeholders

Discussions of CI development have recognized that tensions and cultural differences between domain scientists and developers can be disruptive [8, 10]. 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 how their interests—their "stakes"—influence the design of the CI with which they are associated. 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. For example, the CAMERA project includes system administrators, network administrators, bioinformaticists, software developers, hardware specialists, outreach personnel, and others. Each of these roles brings specific interests to the project; we focus here on a subset of these roles as they relate to the scientists whom we interviewed for our study.

5.1. Domain Scientists and Computer Scientists

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)*

In this case, the domain scientists are training to perform tasks similar to the work of the CAMERA developers. But the development team also has a diverse set of activities and interests that go beyond simply building systems. Because CIs often require cutting edge hardware, software, algorithms, and other computational abilities, computer science researchers are often involved in development processes. These computer scientists have their own research agendas; for them, the CI serves as a test bed or proof of concept for their work. CAMERA computer scientists are involved in research about the design of operating systems for cluster computers, high-speed networking, the design of "middleware" technologies, etc.

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.

5.2. 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)*

This scientist's work requires significant computational capacity. In order to meet that need, he has been relying on computing resources within his own laboratory, and spending his own grant money to develop these systems. CAMERA represents an opportunity for him to shift this burden out of his lab, freeing up valuable staff time and grant money for other purposes.

Scientists might also gain indirect benefits from CAMERA's success. Reputation has been called the "coinage" of science [17]. Merton claims that reputation accrues to scientists through "public recognition of one's scientific contributions by qualified peers" (p. 619). Traditionally, the primary way biologists would earn this recognition is through citation of their publications. DNA sequence databases provide another route for generating reputation. When scientists make important comparisons to a sequence or set of sequences in the database, they often cite the contributors of those data in papers about the new data. Thus, scientists who contribute DNA sequences to the database have a stake in ensuring that the system makes it possible to discover the data they contributed.

One scientist who is designing an analysis tool that may be included as part of the CAMERA system expressed a similar sense of his reputational stake in CAMERA:

> *It's going to be presented as a community tool…. You can publish papers based on that and 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)*

This scientist is working in a metagenomics laboratory, but being involved with the CAMERA development effort will be a route through which he can build his scientific career. His reputation is entwined with the reputation of the CAMERA system.

Database systems like CAMERA also provide one route through which research community values and boundaries are established. Bietz and Lee [18] found that sequence databases act as "boundary negotiating artifacts." The design of the system can favor particular scientific questions and approaches, making it easier not only for the scientist to use those approaches, but also to build a community of researchers around a particular set of scientific questions. The CI becomes a site for enrolling others in the production of science [19]. 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. Thus, 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. But they are concerned with more than just the success of the project itself: CAMERA is part of a larger Marine Microbiology Initiative. For example, the 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's stake in CAMERA's development lies not only in the successful outcome of the project itself, but also in CAMERA's ability to provide useful service to GBMF's other projects that rely on it.

Developers have a direct interest in CAMERA's success because the project is paying all or part of their salaries. But this is rarely their only stake. Some of the developers are looking for opportunities for learning or to be involved with a project that that has a large impact. Others are concerned with advancing their research agenda.

> *Working with these different science projects helps me in figuring out the requirements and checking out how these requirements can actually be fed back into the computer science research schools…. It ties in to my research goals, my career goals. (Jayden)*

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 genomics and metagenomics databases, requiring at least some level of data sharing and compatibility with these other systems [18]. Thus, these other database systems and their users also have an interest in how CAMERA is designed. As an influential system, if CAMERA adopts a particular data standard it encourages others to adopt the same standard, so standards-development bodies also have an interest in CAMERA's development.

5.3. "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 developers rely on several techniques to understand the community. For example, the Scientific Advisory Board (SAB) includes domain scientists who meet regularly and provide guidance for the project.

> *So the initial scientific advisory board for CAMERA was ten scientists that worked in various areas of metagenomics. And so that was really our effort to make certain that CAMERA was hearing from the scientific community.(Andrew)*

In order to develop more specific user requirements, the developers have also adopted a strategy of working with individual scientists who serve as model users. These "early adopters" work closely with developers to figure out the best way to represent and import their data into the CAMERA database. This work is regarded as creating a template for importing future data sets.

Like Ribes & Finholt [12], we find that "community" takes on special significance in CI development. They trace the "formation of a single community from heterogeneous beginnings" in the WATERS project (p. 115). In contrast, we find that in the CAMERA project the use of the term "community" takes on a number of different meanings. We heard participants use "community" to refer to the set of 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)*

Metagenomics is science in action: what Latour describes as emergent, unestablished scientific facts and processes [19]. CAMERA's mission is not simply to make established science more efficient—it is explicitly part of a larger strategy to create new, transformative science. CAMERA participants still use community "as a short-hand for issues of representation" [12], but the term remains usefully vague. The fluidity of the term 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.

5.4. The Evolution of Technology, Research Questions, and Stakeholders

Even though CAMERA is mandated to serve marine metagenomics, this field will only be part of the eventual community of stakeholders.

> *CAMERA started out as a marine microbial ecology sort of research endeavor or research resource. But it doesn't make biological sense to those who would explore the marine world to limit access to only marine data. I mean, microbes are prevalent to everybody. (Michael)*

Other researchers who study microbes from the soil, the human gut, or many other environments are likely to use and contribute to CAMERA. While CAMERA's resources are geared toward metagenomics, they may also be useful to genomicists and geneticists. It is worthwhile to note that, even in this early stage of development, CAMERA plans to expand its user base in terms of size and scope. The result of adding data from other scientific areas would be an inevitable shift in the landscape of stakeholders. By design, the stakeholders will change over time.

Of the variety of stakes and stakeholders that are involved in building infrastructure, some directly relate to the obvious purpose of enabling the creation of scientific "facts." Many of the stakes, however, have their roots in social and organizational structures of scientific practice, a fact that developers recognize.

> *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)*

In order to understand how socio-technical systems like CIs are developed, we must look beyond simple dichotomies in order to grasp the complexities of the problem space of infrastructure creation.

A useful way to understand the evolution of the stakeholder landscape for cutting edge science is to look at the loosely, sometimes barely, connected people that form around research questions. One of the challenges of designing systems for CI environments is that it is insufficient to look at established organizations and functional groups. Communities of interest form around new questions, but it is somewhat of overstatement to even describe these question-driven groups as "communities" because the science is too new. A metagenomics conference draws many scientists who do not self identify as metagenomicists, but are merely interested in metagenomics approaches.

A self-identified metagenomicist, a collaborator in the development of a system for metagenomics, describes three stakeholder communities trying to use a particular database as moving targets:

> *There are at least three moving targets in this project. And that is that there are the ecologist metagenomics people, there are evolution people that are more interested in the evolution of the sequences, you know, what they're telling you about evolution; which is actually quite different how you analyze the data in this case. And then there are just the people that are thinking, like, just genomes and glorified genomes, right. And that's also a very different way of looking at the world. I think that that's a big failing that we didn't recognize that in the beginning as much as we should have…. (Christopher)*

One of the implications of the desire to design a system for people with different scientific inclinations is that, even if the immense amount of work necessary to collect and aggregate data has been successfully accomplished, scientists ultimately need different outputs and tools to achieve them whether it be a genome browser, a statistical package or a fully annotated genome.

Scientists from numerous fields are using these systems and each field brings it own approaches and viewpoints. While the diversity of scientists and scientific interests is challenging, the larger challenge is that research questions are continually evolving.

The instruments will continue to improve. But they're never going to be perfect because we're continuing to push the boundaries. So the kinds of scientific questions we can answer will keep extending. So we'll have demands for new instrumentation. We'll have demands for new software tools… But I think we also feel that we don't know the range of questions fully. And so the same is true for software tools. (Frank)

Scientists will continue to generate new discoveries and technologists will continue to develop new technologies. These discoveries will enable new methods, but will also open up a new range of addressing previously unanswerable research questions. As research questions, data, tools, and practices shift and change so too will the communities around them. Our interviews reveal that, over time, scientists consider themselves to have switched fields and to have adopted new research methods in order to investigate a compelling research question. At least in the case of metagenomics, we see that stakeholders do not map to organizations or functional groups. Rather, stakeholders map loosely but inexactly to disciplines that already have very porous borders.

6. Discussion

Our investigation of CAMERA reveals a slightly different picture than has been reported for other CI projects. Where Spencer, et al. [10] focused on the interactions between domain scientists and developers, and Ribes & Finholt [12] saw community as a central organizing concept, we find that CAMERA development is occurring against a backdrop of multiple stakeholders with a variety of interests. This leads us to reconsider the role of the "user" in design.

While Mackay [20] describes configuring the user as bidirectional, configuring infrastructure is multi-directional. Developing an infrastructure requires an even greater complexity of actions and configurations. Given the complexity of a large-scale, long-term development project such as that for CI, the notion of a user that can be configured by "someone" or even by a single organization may be too simplistic a model for understanding the work of infrastructure building. Information infrastructures result from incremental restructuring of sociotechnical relationships.

In CI development, it is useful to reframe "eliciting user requirements" as a process of discovering and managing a complex set of continuously evolving stakeholder 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 [21]. The selection of stakeholders and stakes will depend on factors ranging from how the project is funded to the history of the scientific domain. 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.

In order to support scientific collaboration it is useful to map out scientific stakeholders according to scientific questions, and not according to domain or institutional allegiances. Similar to designing for other types of organizations [22], the mapping of concerns must be done iteratively to keep pace with change. In discussing the Hospital Information Systems (HIS), Balka et al. [23] discuss three nested levels, also called arenas, of system and workplace design context, institutional/organization context, and political and policy-making context (see Figure 1).

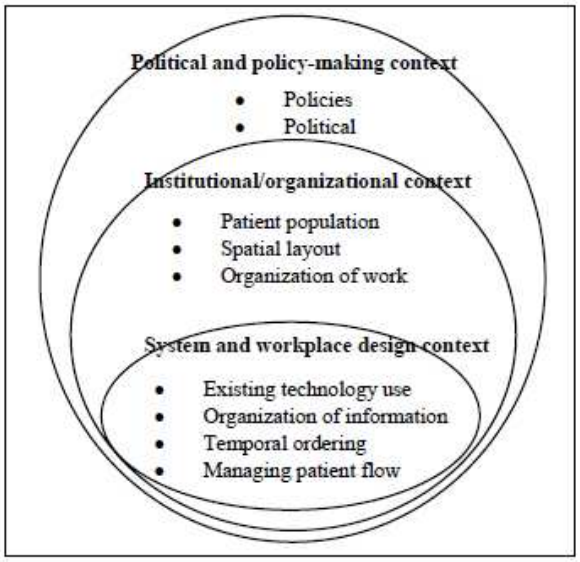

Figure 1: Typology of Health Information Systems (from [23])

Balka et al.'s [23] excellent model suggests several interconnected and interdependent variables that illuminate the context of action. However, there is key difference between the scientific milieu as represented by our participants: the system and work are geographically distributed and the system is also organizationally dispersed. In the hospital context there are departments with different priorities and practices, but they can be nested within one institutional or organizational context, and that organization can be nested within one highly complex policy context. In the case of our participants however, while the system will be slowly evolving, the workplace context around it is variable to the point

where not only workplaces changing, but new workplaces are gradually coming on board and others are leaving. To compare Balka et al.'s model to the metagenomics context, we can replace departments with laboratories, and we can replace the hospitals with universities or research centers. Certainly universities and research centers exist within their own policy and political contexts. But at this point the mapping of HIS to information systems for cutting edge science breaks down. For example, the ambitious scientists who are conducting the most innovative (and risky) types of research advance their careers by getting peer-reviewed publications accepted and published. Our interview data show that while scientists are accountable to their organizations, they are also typically subject to a very diverse set of commitments. One scientist told us about his laboratory:

> *[Marine Research Organization] was formed by the National Environmental Research Council in the UK.... Now we're open to grants from any research body or any funding body, I should say, and we can also perform commercial applications. We have a commercial wing... which attempts to commercialize some of the research ideas that we develop and look for commercial opportunity externally.... The students I have working in the lab, they're based at different universities around the country for their degrees. They're registered there, but they spend 100 percent of their time with us.... I also share a collaborative supervisory role for a PhD student in [a United States university laboratory].... I work on a relatively diverse array of projects. (Thomas)*

This scientist is accountable to his own organization, his governmental and commercial funders, and the various universities where his students are enrolled. He is also accountable to his peers conducting similar and competing research, to labs that are helping to provide data in exchange for help with analysis, and to other collaborators who are providing expertise in exchange for access to data.

There are two things to note in this example. One is that the workplace design context is frequently not unitary, but can be dual, tripartite, and so on. The base level, then, is imperfectly bounded by "the laboratory" because so much research (even at the granular level of a single research paper) is interdisciplinary and spans workplaces. The organizational context similarly multiplies not only in number but also in kind. The institutional imperatives of a for-profit research center differ substantially from that of a non-profit research center. Similarly the institutional imperatives of a shotgun sequencing center is quite different from those of a research laboratory and from a consortium of people establishing database standards, or an NSF (US National Science Foundation) program officer looking for innovative research to fund. Yet all of these parties, and more, must be involved in the development of a single cyberinfrastructure.

The geographic split between the system itself and the work context provides an additional challenge. A system developed and maintained at UC San Diego with input from some distributed collaborators is used in a variety of workplace and organizational contexts. The system is used in multiple workplaces each embedded in different organizations. While workplace contexts (e.g. a lab that belongs to the a university organization) do sometimes nest within organizational contexts at other times organizational contexts merely overlap workplace contexts (e.g. a lab may have a project funded by NSF, but this does not mean that all the work taking place in the lab or even with a given system is beholding to the rules of NSF). Similarly, organizational contexts may sometimes nest neatly within policy contexts, but in the cyberinfrastructure realm there are many policy contexts that do not apply directly ot the organization, but rather some subset of work. The large-scale, usually interdisciplinary research undertaken with and through cyberinfrastructure systems presents an extremely diffuse model of work, organizations, and policy. Perhaps taking a few dozen instances of Balka et al.'s nesting model, and scattering them across a canvas where multiple models as depicted above may intersect at any one, two, or three of the levels would more accurately depict the range of stakeholders.

It is worth underlining that "the system" is not unitary. Many of the laboratories that use centralized CI systems also develop their own tools. These tools may be considered outside of the system, but our interviews indicate that these bespoke tools are crucial for the accomplishment of work that prompted use of the system in the first place.

Taking a more holistic view of complex information systems and minding the interaction of different elements at the levels of workplace, institution/organization, and policy is a very important step forward. In order to understand the design of cyberinfrastructure systems, we need to broaden our understanding of stakeholders beyond simple dualities, build on more sophisticated models [23], and begin the difficult work of modeling more complex stakeholder interactions between networks of workplaces and organizations.

7. Conclusion

As collaboration becomes more complex, so must our models of work. CI development should be seen as stakeholder-driven; when scoping the design space for an infrastructure, a broad array of stakeholders should be considered and involved. Whereas participatory design methods regard broad involvement of all stakeholders as desirable, the broad involvement of stakeholders is essential to the successful creation of CI. To be on the cutting edge of science is to cultivate an infrastructure that, by necessity, is always changing. 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.

Systems-based design approaches that place stakeholders on the outside of design considerations are inappropriate for the design of large-scale information infrastructures. Efforts to develop for a broad array of stakeholders or for users who interact directly with the system are often talked about as separate endeavors, but our research shows that these endeavors often overlap, blend, and at times become indistinguishable. Some stakeholder-users may interact only superficially with the system, whereas some stakeholders who never interact directly with the system are heavily invested in the success of a system.

Cyberinfrastructure systems are geographically distributed, designed to support the investigation of evolving practices, evolving specific research goals, and are not organizations in the traditional sense of the word. Researchers involved in the investigation and development of collaborative technologies that comprise and support large scale scientific cyberinfrastructures will need to explore and find ways to model the complexity of a milieu where the landscape of activity is perhaps better described as a loosely connected network of stakeholder concerns that smash apart the utility of simple dichotomies and that also do not map to somewhat neatly nested levels.

The most useful way to organize the rats nest of stakeholders, who are at different levels that may or may not nest, into a comprehensible—and therefore researchable and supportable—alignment is to focus not only on the system under construction, but to follow the threads of a specific research area. Research areas, too are not unitary, but similar types of research questions entail certain types of data, metadata, levels of analysis, supporting infrastructure, and analytical tools [18]. In this way, we can establish a class of workplace activities that are quite similar despite being distributed across organizations, and that often have some similar political concerns (e.g. trying to get funding from the same NSF program), or policy concerns (e.g. fulfilling requirements to make data available to the national genomics data base), even though many elements of workplace, organization, and policy are different. As research questions and research areas fall in and out of favor, so too will the stakeholders involved. How to model and support a context of action in which the stakeholders involved are in constant flux is a thorny challenge for future research in cyberinfrastructure and other complex, distributed, large-scale systems.

8. 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).

9. References

[1] W. J. Orlikowski, "Learning from Notes: organizational issues in groupware implementation," in *Computer Supported Cooperative Work (CSCW)*, Toronto, Canada, 1992, pp. 362-369.

[2] F. Berman, "The human side of cyberinfrastructure," *EnVision*, vol. 17, p. 1, 2001.

[3] C. P. Lee, P. Dourish, and G. Mark, "The human infrastructure of cyberinfrastructure," in *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work* New York: ACM, 2006, pp. 483 - 492.

[4] B. Friedman, P. H. Kahn, J. Hagman, R. L. Severson, and B. Gill, "The Watcher and the Watched: Social Judgments About Privacy in a Public Place," *Human-Computer Interaction*, vol. 21, pp. 235 - 272, 2006.

[5] 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.

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

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

[8] 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.

[9] D. E. Atkins, K. K. Droegemeier, S. I. Feldman, H. Garcia-Molina, M. L. Klein, P. Messina, D. G. Messerschmitt, J. P. Ostriker, and M. H. Wright, "Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure," National Science Foundation, Washington, D.C.2003.

[10] 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.

[11] J. P. Birnholtz and T. A. Finholt, "Cultural challenges to leadership in cyberinfrastructure development," in *Leadership at a Distance*, S. Weisband, Ed. New York: Lawrence Erlbaum Associates, 2007, pp. 195-207.

[12] 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.

[13] 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.

[14] R. Seshadri, S. A. Kravitz, L. Smarr, P. Gilna, and M. Frazier, "CAMERA: A community resource for metagenomics," *PLoS Biology*, vol. 5, p. e75, 2007.

[15] M. Brown, L. Smarr, T. DeFanti, J. Leigh, M. Ellisman, and P. Papadopoulos, "The OptIPuter: A national and global-scale cyberinfrastructure for enabling LambdaGrid computing," in *TeraGrid '06 Conference*, 2006.

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

[17] R. K. Merton, "The Matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property," *Isis*, vol. 79, pp. 606-623, 1988.

[18] 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.

[19] B. Latour, *Science in Action*. Cambridge, MA: Harvard University Press, 1987.

[20] H. Mackay, C. Carne, P. Beynon-Davies, and D. Tudhope, "Reconfiguring the user: Using Rapid Application Development," *Social Studies of Science*, vol. 30, pp. 737-57, October 2000.

[21] N. Bos, A. Zimmerman, J. S. Olson, J. Yew, J. Yerkie, E. Dahl, D. Cooney, and G. M. Olson, "From shared databases to communities of practice: A taxonomy of collaboratories," in *Scientific Collaboration on the Internet*, G. M. Olson, A. Zimmerman, and N. Bos, Eds. Cambridge, MA: MIT Press, 2008, pp. 53-72.

[22] J. E. Bardram, "Organisational Prototyping: Adopting CSCW Applications in Organisations," *Scandinavian Journal of Information Systems*, vol. 8, pp. 69-88, 1996.

[23] E. Balka, P. Bjorn, and I. Wagner, "Steps toward a typology for health informatics," in *Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work* New York: ACM, 2008, pp. 515-524.