Sustaining the Development of Cyberinfrastructure: An Organization Adapting to Change

Matthew J. Bietz
Department of Informatics
University of California, Irvine
Bren Hall 5042, Irvine, CA 92697, USA
mbietz@uci.edu

Toni Ferro, Charlotte P. Lee
Human Centered Design & Engineering
University of Washington
423 Sieg Hall, Seattle, WA 98195, USA
{tdferro, cplee}@uw.edu

ABSTRACT

Cyberinfrastructures are virtual organizations comprised of people and large-scale scientific computational infrastructures. Cyberinfrastructures endeavor to support "cutting-edge" science and must continually evolve and be under development in order to maintain their relevance and usefulness. This qualitative study of a cyberinfrastructure development project to support the new science of metagenomics investigates how sustaining cyberinfrastructure entails continually realigning the relationships among people, technologies, and organizations.

Author Keywords
Cyberinfrastructure, e-Science, Sustainability, Synergizing, Infrastructure, Organizations.

ACM Classification Keywords
H.5.3. Group and Organization Interfaces.

General Terms
Design; Management.

INTRODUCTION

Cyberinfrastructures (CIs) are virtual organizations comprised of people and large-scale scientific computational and networking infrastructures. In order to answer pressing scientific questions in areas like human disease and global warming, scientists are engaging in large-scale distributed and interdisciplinary collaboration around very large data sets from a wide variety of sources. New infrastructures are needed to meet the communication and computation demands of contemporary collaborative science.

There is an emphasis on sharing tools and data in order to enable new forms of collaborative scientific production while making the most efficient use of available resources. However, as the scale of science grows to include larger aggregated data sets and to depend on greater levels of computational power, the question arises of how to ensure that the infrastructures that support this digital scientific knowledge base will be preserved and maintained over time. Long-term sustainability of systems and resources has become a key concern for large-scale scientific computing (a.k.a. cyberinfrastructure, e-Science, scientific Grids, etc.).

Building on the work of Star and Ruhleder [27,29], we adopt a relational view of infrastructure. Infrastructure is less a thing in and of itself than a set of relationships among people and technologies. When we refer to CI, we are referring to arrangements of technologies, individuals and organizations [15]. This research explores and reframes the sustainability of cyberinfrastructure as an ongoing process of relational maintenance.

Our research site is the Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA). This project is developing advanced computational resources to support and promote the emerging community of scientists working in the new field of metagenomics. In this paper, we investigate how the developers of CAMERA understand, manage, and respond to change, and we use these findings to understand relational maintenance as a set of strategies necessary for accomplishing infrastructure.

LITERATURE REVIEW

Cyberinfrastructure

Trends toward big science and big data [2,10] are increasingly pushing scientists to collaborate across traditional organizational, geographical, temporal and disciplinary boundaries. The development of large-scale cyberinfrastructures is part of this trend. A US National Science Foundation (NSF) Blue Ribbon Panel defined cyberinfrastructure as follows: "Cyberinfrastructure refers to infrastructure based upon distributed computer, information and communication technology. If infrastructure is required for an industrial economy, then we could say that cyberinfrastructure is required for a knowledge economy" [3]. Spurred on in part by the findings of this panel, there have been significant investments made by the NSF and other agencies in developing cyberinfrastructure.

Cyberinfrastructure is more than software and hardware. It also encompasses the human infrastructure of CI, which in

CSCW'12, February 11–15, 2012, Seattle, Washington, USA.
Copyright 2012 ACM 978-1-4503-1086-4/12/02...$10.00.

large projects is “a vast series of overlapping traditional organizations, consortiums, loosely organized groups, and networks” [15]. In this paper our concern is with understanding the processes through which infrastructural relationships are created and sustained [13,28]. We draw heavily on the work of Star et al. [27,28,29] that understands infrastructures to be fundamentally relational. Infrastructures are embedded inside other structures and technologies, function beyond the scope of a single event or site, and are closely linked to communities of practice [29]. Our goal is to recognize, understand, and ultimately better support the collaborative work of infrastructuring.

Cyberinfrastructure and Organizational Sustainability

Cyberinfrastructure is becoming increasingly important for the conduct of science, but many early CI projects were developed as short-term endeavors without much consideration of what would happen to the infrastructural resources when the projects concluded. As more science depends on cyberinfrastructure and early CI development projects are coming to an end, the sustainability of CI has become a pressing issue [24].

Sustainability can be a difficult concept to pin down, however. To sustain cyberinfrastructure means to make it last or endure for long (or even indefinite) time periods. But this leaves open questions about what exactly should be preserved and how to choose among potentially competing infrastructure priorities like sustainability, innovation, growth, and robustness. Our goal here is to inform this conversation by studying and theorizing the work of maintaining cyberinfrastructure. The discussion of CI sustainability has tended to focus in two areas: sustainable institutional resources and preservation of scientific products.

In terms of the issues of sustainable institutional resources, researchers have discussed how to continue to fund cyberinfrastructure given the challenges of long-term planning and resource allocation. Organizations like the National Science Foundation that have funded much of the early development of cyberinfrastructure are poorly placed to provide long-term, ongoing support. Concerns in this area have focused on issues like the reusability of code and components to ensure efficient development and deployment of technological resources [26]. Another set of concerns revolves around incorporating cyberinfrastructures into existing institutional structures like university libraries or IT departments as a way to transition from short-term development projects to a more sustainable funding model [19].

The second factor driving the push for sustainability is a concern that we may be losing valuable (and sometimes irreplaceable) knowledge if scientific outputs are not adequately preserved [13]. In this era of “big data” research, datasets are seen as knowledge products in and of themselves, and it is important to ensure long-term storage and curation of the data [18]. Similarly, as models and algorithms are becoming an increasingly central part of the scientific method, there is a push to save research-related software as well [30]. The preservation of scientific artifacts would ideally allow future scientists to judge the validity and accuracy of scientific conclusions. Just as important, however, is that data and codes must be preserved for potential reuse by future scientists [16].

This paper advances our understanding of a third aspect of sustainability, specifically, how to maintain the persistent human and technological arrangements that comprise cyberinfrastructure. Ribes and Finholt discuss the dilemma of building a sustainable infrastructure in the world of software development, which is known for rapidly advancing hardware, development platforms, and programming languages [23]. They see cyberinfrastructure as a sustainable, human-technical collective and articulate a set of tensions which infrastructure developers face when thinking about long-term sustainability. For example, the “development vs. maintenance” tension arises from the necessity of doing ongoing upkeep work even though developing new tools and resources tends to be more valued and rewarded. These tensions describe the pressures that drive infrastructure development. This framework has much in common with the approach taken here, especially in its recognition that developing cyberinfrastructure “requires creative attention to issues of sustainable technology, persistent human arrangements, and institutional resources” (p. 379). This paper builds on this observation that sustainability is needed, and investigates the work of a CI development team to demonstrate how sustainability can be achieved.

Developing Cyberinfrastructure Through Synergizing, Leveraging, and Aligning

Recognizing the challenges and necessity of sustaining cyberinfrastructure, CSCW research has turned to investigating how sustainability is accomplished. In order to investigate organizational sustainability and organizational change in CI, Bietz, et al. [6] introduce the concept of *synergizing*. Synergizing recognizes that infrastructure is necessarily woven into existing structure, social arrangements and technologies” [29]. Synergizing is “the work that developers of infrastructure do to build and maintain productive relationships among people, organizations, and technologies.” Synergizing specifically draws on the concept of synergy, the increased effectiveness resulting from combined action or cooperation [22]. This systems-level, relational view of infrastructure work provides the framing for our discussion of cyberinfrastructure sustainability.

Synergizing encompasses two specific kinds of work activities: leveraging and aligning. Leveraging is the work a development team does when it uses its existing relationships (between people, organizations, or technologies) to build or strengthen other productive relationships. For example, a development team may leverage connections within its own university to gain access to cutting edge technology or other resources. In addition, leveraging an existing network can result in less alignment work (because an existing relationship indicates that some level of alignment work has been done previously).

Aligning is the work that developers do to make relationships among people, organizations and technologies productive and functional within a specific CI. An example of alignment work is developing an application programming interface (API) that allows two component technologies to communicate with each other. However, alignment work is not limited to creating connections between technologies and includes the work required for individuals and organizations to be able to collaborate. Alignment work can include the work of developing data sharing policies that are amenable to all participants, or ensuring that technological security arrangements are in line with those policies. Alignment work is "the work necessary to create enough compatibility between entities so that the relationship can be productive."

Synergizing is distinct from articulation work [11,25] because it ensures that a common field of work exists, whereas articulation ensures that work goes well and that complexity is controlled within an existing field of work. Synergizing creates the situation in which articulation work can be enacted. However, both synergizing and articulation work are fundamentally concerned with the work done to make other work possible.

Relationships among human and technological components (like the components themselves) change over time, and must be maintained and tended if they are to continue to exist. Synergizing affords a consideration of cyberinfrastructure sustainability that concentrates on how these relational structures are managed over time.

SITE AND METHODS

The Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) is a cyberinfrastructure development project intended to provide resources for high-volume data storage and analysis in metagenomics. Metagenomics is a "new science" aimed at understanding the genomic characteristics of microbial populations [21]. New laboratory technologies, in combination with advanced computational capabilities, allow scientists to change the level of analysis from the single organism (as in traditional genomics) to entire populations of microorganisms. For example, the Human Microbiome Project [32] is focused on characterizing the bacteria and other microbes that live in the human body in order to understand how these microorganisms contribute to human health and disease.

Metagenomics depends on access to large databases of genomic data and significant computational and networking power [5]. Microbiologists in our study report a shift away from traditional hands on research in the field or laboratory so that now more than 90% of their research is conducted "*in silico*." Genetic sequencers produce vast amounts of data about each biological sample, which scientists analyze with a variety of statistical techniques, many of which require comparing the new sample to large databases of reference genomes. While other genetic and genomic databases exist, scientists were concerned that none were capable of supporting the specific data and analysis demands of this new science. In response to this concern, the Gordon and Betty Moore Foundation (http://www.gbmf.org) commissioned the development of CAMERA. The foundation funded the California Institute for Telecommunications and Information Technology (Calit2) at the University of California San Diego (UCSD) and the J. Craig Venter Institute (JCVI) to build a new cyberinfrastructure for metagenomics. The project began in 2006 and was funded at $24.5 million over seven years.

Data Collection and Analysis

The CAMERA cyberinfrastructure development project was the site for the current research. We engaged in two extended periods of investigation with the CAMERA project team. In both periods, our investigation included a mix of interviews with project participants and observation of everyday activities and meetings. We will refer to these two periods as "Round 1" and "Round 2."

Round 1 began in the late summer of 2007, but most observations took place between January and May of 2008. During that 4-month period, the lead author attended weekly project meetings, ad hoc meetings, and spent at least one day per week working from an assigned desk in the development team area. We interviewed as many members of the development team as we could, some of them multiple times.

Round 2 of our investigation took place from May through September 2010. Again, the lead author attended various project meetings, observed daily work (although there was no assigned desk in Round 2), and interviewed as many members of the development team as would participate.

These two time periods were chosen to be approximately two years apart and to fit with both the researchers' and participants' schedules, but there were no other theoretical motivations for choosing these specific dates. Our goal was to develop two snapshots in the life of the project that could serve as a basis for comparison and help us understand how projects may change over time. It should also be noted that this time period is relatively short in the life of infrastructures. Ribes and Finholt [23] borrow the concept of the "long now" to discuss issues of the long-term in infrastructure development. Our three-year study period gives us some perspective on change, but would be better described as the short- to medium-term. This time period does not allow us to comprehensively catalog all of the potential aspects of long-term change.

Even so, investigation at this scale covers substantial organizational change and yields important characteristics of the work of cyberinfrastructure sustainability. In late

2008, the CAMERA development team underwent a significant reorganization which also included a major staff shakeup. When we conducted Round 2 of our investigation, one of the earlier PIs was no longer on the project, and of the development team professional staff, only three individuals were still working on the project. Coupled with the rapid pace of change in the developing science of metagenomics, we saw a number of interesting changes between the two rounds of investigation.

While our primary investigations involve those individuals and organizations who are directly involved in the work of development, we take a broad view of the site and include a wide swath of stakeholders including funders, the "users" of the system, project collaborators and competitors, and members of broader scientific communities who hold a stake in the development of this and similar cyberinfrastructures.

During the two-year period when we were not on site with the CAMERA team, our focus shifted to developing a broader understanding of the landscape of cyberinfrastructure for metagenomics research. We interviewed microbiologists, bioinformaticists, computer scientists, and representatives of funding agencies. We interviewed both users and developers of several major genomics and metagenomics databases. We attended conferences and workshops devoted to metagenomics research, database development, and the development of standards for genomic data and metadata. Over a period of five months, one of the authors attended weekly laboratory meetings at an academic molecular biology laboratory engaged in metagenomics research.

In all, we conducted 43 semi-structured interviews with 33 CAMERA team members, other stakeholders, and scientists in the metagenomics field. Interviews lasted from 20 to 102 minutes (median: 52 minutes). We attended 30 regularly scheduled CAMERA meetings, numerous ad-hoc meetings, and observed the ongoing work of the CAMERA development team, resulting in over 100 hours of observation. Transcripts of interviews and field notes were analyzed using grounded theory techniques and Atlas.ti qualitative data analysis software [4,7,12]. All participants are identified by pseudonyms. Data were open coded as they were generated. Coding continued as new data came in, both building on and extending the existing code list. Descriptive and analytic memos were written during ongoing analysis of the data. Coding and memoing were iterative: after writing memos, we would return to the texts to further enhance and refine the coding scheme. Through this process, we developed the set of themes that we discuss in this paper.

FINDINGS

Like others who have studied sustainability, we observed that a state of ongoing change is the norm for cyberinfrastructure. Sustainability cannot be a matter of maintaining the status quo. Only by responding to new organizational arrangements, technologies and scientific needs can an infrastructure remain useful. In this section, we use the analytical lens of synergizing to explore in detail how cyberinfrastructure responds and adapts to ongoing change. Cyberinfrastructure development is a process of creating and maintaining relationships among people, organizations and technologies. Documenting synergistic activity is difficult because the complexity of the systems requires an analytic complexity that cannot be captured by a focus on a single analytical unit. The findings below take only elements of cyberinfrastructure—human infrastructure, technology, and science—and trace how those elements interact and co-evolve with other infrastructural elements in order to illustrate the work of how organizational stability is accomplished.

Human Infrastructure: Changing and Stabilizing

Understanding sustainability through the lens of synergizing requires a consideration of how infrastructures adapt. The notion of human infrastructure underlines the importance of recognizing the multitude of collaborative forms (organizations, networks, teams, etc.) necessary for accomplishing infrastructural work. Therefore in order to understand how the sustainability of cyberinfrastructure is managed, we explore how those entities interact.

When we returned to the CAMERA project for Round 2 of our study, the most obvious change was that the project had undergone a significant shift in the human infrastructure. This began as a change in the organizational alignments of the project. CAMERA was originally a partnership of Calit2 and the J. Craig Venter Institute (JCVI). Calit2 is a research institute of the University of California, and the CAMERA team was located entirely on the UC San Diego campus. JCVI is an independent genomic sciences institute with most of its staff in Maryland, USA. During Round 1 of our study, JCVI had decided to leave the CAMERA project, and interactions with individuals at that institution were becoming less frequent. JCVI had provided a good deal of biological and technological expertise, and there was a concern that their leaving could represent a loss of capability. CAMERA's leadership felt that the project needed to quickly partner with another organization to deliver on the promises made to the funders and the scientific community. As our Round 1 engagement was wrapping up, the project leaders had decided to partner with the Center for Research on Biological Systems (CRBS) in order to benefit from their experience with other similar projects. CRBS is another research unit at UCSD with significant experience developing cyberinfrastructures, including the Neuroscience Information Framework, the National Biomedical Computation Resource, and the Biomedical Informatics Research Network. While CRBS and Calit2 are independent units, they have become closer collaborators, especially since they began working together on CAMERA. In Round 2, even though Calit2 was still involved, CRBS was clearly the lead organization for the development work, with key project leadership positions occupied by CRBS-affiliated staff.

This change in organizational arrangements had also resulted in significant staff turnover. Of the Round 2 staff, only two of three members of the Executive Committee, one of six members of the Leadership Team, and three of twenty-five professional staff had been on the project during Round 1. Many of the staff working on CAMERA in Round 2 had previously worked with CRBS on other cyberinfrastructure projects.

The synergizing lens, which focuses attention on the work of building and maintaining infrastructural relationships, helps us understand how this transition was managed. The decision to collaborate with CRBS was seen as a way to sustain infrastructural capabilities after it became clear that JCVI was going to leave the project. However, the alignment work necessary to make the new relationship with CRBS productive would take time. A CRBS staff member on the Round 2 CAMERA team explained how CRBS took over support of JCVI-developed applications:

> *"We had a phased transition plan… where our teams worked together with their teams, and that was to gain an understanding of the system…. And then after a certain period of time we took over the primary support for the components for the entire system…. So, it wasn't an abrupt cutoff." (Johann, CRBS staff, Round 2)*

Switching partnerships from JCVI to CRBS was not an instantaneous change, but was managed in phases so as to minimize disruption to the infrastructure. This transition in the human infrastructure involved first bringing in the new organization (CRBS) and personnel and working to make those new relationships productive before JCVI transitioned out of the project. We also found that while only a few of the professional staff were still on the project, those staff had assumed larger responsibilities in the project, and they played important roles in providing historical information to the new team members.

The transition was easier than it might have been because CRBS and the new participants were already embedded in many of the same structures as Calit2 and the Round 1 CAMERA staff. CRBS was also based at UCSD, had been involved in cyberinfrastructure development for years, and had worked closely with scientists in biology and related fields. This simplified the alignment work in a number of areas. For example, mechanisms for moving money between Calit2 and CRBS were already in place because they were both part of the UCSD financial systems, and CRBS was already familiar with the UCSD computer networks and other technological resources on campus. Synergizing highlights that being embedded in multiple, overlapping networks and systems is an important resource for cyberinfrastructure development, as existing infrastructural relationships are *leveraged* to create new relationships.

To understand how CAMERA was sustained through this organizational change, it is also important to recognize that the human infrastructure of cyberinfrastructure is larger than the development team. Even as these changes were going on within the project staff, other stakeholder groups remained relatively constant. CAMERA's 10-person scientific advisory board played a significant role in setting the project's vision and priorities, and between Rounds 1 and 2, only two people left and one new person joined this board. In both rounds, we saw that not only did the project funding come from the same source, the same program officer was involved. And even though the development team was almost completely different, they were still working with many of the same scientists and the same laboratories in Round 2.

CAMERA's ability to serve the metagenomics community was preserved, even as organizational arrangements and development staff shifted. We see this stability resulting from a variety of factors, including a phased transition, leveraging other infrastructural relationships, and drawing on the stability of the project's other stakeholders. Maintaining this stability of the whole even as various aspects change is an important part of cyberinfrastructure sustainability.

Changing Technologies: Causes and Consequences

The relational embeddedness of cyberinfrastructure implies a degree of dependency or contingency among infrastructural components. Changes in one area lead to changes and adaptations in others. For example, changes in human infrastructure, precipitated changes in technology, which in turn precipitated still other changes in human infrastructure and technology.

We found that certain changes in the technological infrastructure of CAMERA were a direct result of the reorganization of the human infrastructure. During Round 2 of our study, the CAMERA team released version 2 of the CAMERA resource, the first major release since CRBS had joined the project. One participant discussed the tight coupling between the human infrastructure and the new technologies:

> *"I mean with the changeover to CRBS it's not - I wouldn't say it's just a management change. With the change to CRBS completely changed the site itself." (Martin, Professional Staff, Round 2)*

After CRBS joined the project, CAMERA adopted a set of technologies that included portal technology for the web site, a different database platform, and various project management and development tools. While some of these technologies provided new or improved functionality, many replaced existing similar technologies. However, the new team was already familiar with the new technologies, and they were used widely in the CRBS organization. By adopting these technologies, CAMERA could benefit by leveraging a number of other relationships: the team would gain access to a wider pool of expertise from other CRBS

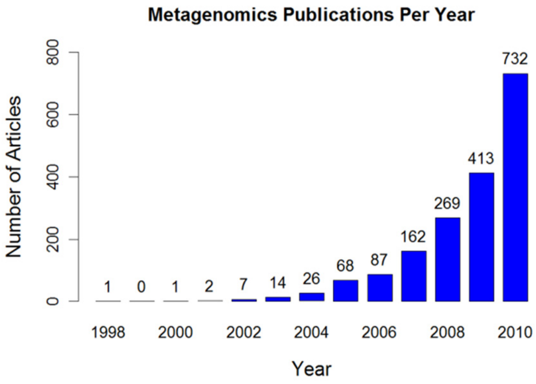

Figure 1: Number of publications per year returned from a PubMed search for "metagenom*" (April 30, 2011).

projects, they could shift some of the burden of system administration outside of CAMERA and onto the CRBS infrastructure, and they could participate in CRBS license agreements for commercial technologies.

However, adopting the new technologies also required making adjustments elsewhere in the system. Pieces of the JCVI code had to be rewritten to interface with new database software. New instructions had to be written to help users manage their accounts on the system. These changes fed back into the human infrastructure, with roles in the projects shifting slightly to accommodate the new technology. This spreading out of actions and responses through infrastructural relationships can be a path to innovation in infrastructure, and in fact, this coupling of infrastructural components is often part of the reason for initiating changes (e.g. making a technology change may draw in a new stakeholder community that needs the new capabilities). In order to realize this innovation and sustain the infrastructure, the CAMERA developers had to engage in an ongoing process of adjusting and tweaking relationships between components to keep them in alignment.

Changing Science

In order to sustain the CAMERA cyberinfrastructure, developers had to adapt to changes in the science and communities they were serving. In the previous sections we discussed the work that developers do to keep the human and technological structures in alignment within the CAMERA cyberinfrastructure. However, in order for the CI to maintain its value to the community of users, it is necessary to maintain relationships that may be outside of CAMERA's control. This work involves similar synergizing processes, but with an emphasis on managing relationships with other systems and infrastructures.

To some extent, the need to adapt to external changes is driven by the development of the scientific domain. During the period in which we were studying CAMERA, we observed a number of changes in the scientific context. 2008-2010 was a period of significant growth for this scientific community, as can be seen in the number of published metagenomics papers per year (see Figure 1). When we started working on this project, metagenomics was still on the fringes of microbiology, and some of our scientists reported that it was difficult to convince journal reviewers that metagenomic studies were valid. By 2010, metagenomics was on a much more established footing.

The growth of the metagenomics community and the development of metagenomics science created another set of sustainability-related pressures on the CAMERA cyberinfrastructure. At the same time that some scientific controversies were closed, others opened. Research questions and methods in the community shifted. In order to stay current and continue to provide useful services, the CI had to adjust to what was going on around it.

For example, in Round 1, the developers were dealing with what contextual metadata—data about the environment from which the metagenomic samples were collected—should be stored in the CAMERA database. One of the key differences between metagenomics data and data from genomic studies is the importance of metadata. While this is a simplification, traditional genomics research has focused on the genomes (hereditary information encoded in DNA and RNA) of individual organisms or species. Metagenomics, on the other hand, uses genetic material to understand the relationships between populations of microorganisms and their environments. For example, metagenomic studies have investigated how populations of organisms are affected by acid mine drainage [1], and whether particular communities of organisms in the gut are related to obesity [31]. In order to understand these phenomena, it is necessary not only to store the genetic sequence information, but also to store information about the environment, e.g. the pH of the mine runoff, or the weight of the study participants. This metadata is also crucial for data aggregation and recombination. Several scientists in our study expressed a hope that soon it would be possible, using a database like CAMERA, to combine datasets in new ways to answer new research questions. For example, even though the two studies mentioned above are looking at very different 'environments,' the data from both may be useful to discover whether all microorganisms that live in acidic environments share certain characteristics. This kind of aggregation across studies requires that all studies in the database share a consistent set of standardized metadata. However, because the metadata are closely tied to particular research questions and disciplinary approaches, deciding which metadata are included in the database can become a contentious negotiation [5].

In Round 1, CAMERA was struggling to decide which metadata fields should be in their database. One participant stated,

> *"Another challenge is the – there are really no standards associated with the metadata* per se*....*

> *There's no sort of well-defined, organized checklist… that a scientist in the field would use as part of the data collection process. So – and that's ongoing and the challenge is to sort of work with the community and other practitioners to generate those standards." (Philip, Executive Committee, Round 1)*

At this time, there was basically no agreement in the community about what metadata should be stored. However, by Round 2, this question was no longer a pressing concern. The Genomics Standards Consortium (GSC, http://www.gensc.org), a group of stakeholders from across the genomics and metagenomics communities published a standard metadata checklist in mid-2008 [8]. While there were still ongoing metadata negotiations, these discussions now took place under GSC auspices. The publication of the standard checklist allowed CAMERA to treat the controversy as closed (at least until the next version was published). The CAMERA development team quickly adopted the GSC checklist: "In the data query application, we follow the standard very strictly" (Alvin, Professional Staff, Round 2). The publication of the standard created a pressure on CAMERA to adapt, but the standard also signaled a level of agreement within the scientific community and allowed CAMERA to shift its attentions elsewhere.

At the same time that some issues were becoming less salient, others were becoming more problematic for the developers. One of these revolved around how CAMERA would support analysis of metagenomic data. In Round 1, the developers were focusing primarily on analyses that would be widely useful to metagenomics researchers. Users could choose from a predefined set of analyses to run against either datasets already in the CAMERA system or datasets that they had uploaded. There were plans to bring community-created analyses into the CAMERA system. The developers believed that software created by end users would be more useful to other community members.

> *"Software coming in from the community is absolutely, absolutely a priority for me. That is something we are working on quite hard.... The software being developed externally is by biologists with a little bit of skill in computer science that know exactly what they're doing scientifically and by definition there's a much higher probability that what they're doing is of interest to the scientific community." (Morgan, Leadership Team, Round 1)*

Even though these scientists may not have significant skill or experience in computer science or software engineering, they are developing software specifically to fill a scientific need that they have experienced. Working with members of the scientific community, the CAMERA team would identify the software that could be useful more broadly. They would then bring this software into the CAMERA systems so that other scientists could use it to run their own analyses.

When we came back for Round 2 of investigation, we discovered that incorporating user-generated analysis software was a much larger part of the project. However, the model for doing so was significantly different and was a response to new pressures from the scientific community. Metagenomic datasets were getting larger and stretching the capacity of in-lab computing resources. The breadth of metagenomics approaches and research questions was also increasing, so that it was less likely that a small number of pre-selected analyses would serve the entire community. The focus shifted from providing a few basic analysis tools to being a more open analysis platform for metagenomics research.

This change in approach involved incorporating a new technology. Whereas before each analysis used its own custom script and operated independently of other analyses, now analyses were conducted using a "scientific workflow" system called Kepler [17]. Kepler allows scientists to create graphical flowchart-like versions of their analyses. These workflows can also be loaded and run on the CAMERA systems, allowing scientists to relatively easily run their own analyses on high-performance computers. Like the community software concept that we saw in Round 1, the workflows create a framework for bringing scientist-generated software into CAMERA. However, in Round 2 the emphasis is on allowing scientists to run their own individual analyses using the CAMERA data and computational resources without necessarily sharing their scripts or having to interact with the CAMERA developers. If scientists do want to make their workflows public, it is simply a matter of changing the sharing settings in their CAMERA account.

The workflow system was a way to respond to the changing needs of the scientific community the cyberinfrastructure was serving. The community was becoming more diverse in terms of its analysis requirements, and the workflows allow scientists to conduct analyses on the CAMERA system that are specifically tailored to their needs. In this case, the work needed to sustain the CI so that it would remain valuable to the labs included leveraging an existing technology (Kepler) which had been available for some time and had a robust open-source academic development community. As a way to help the alignment work of integrating Kepler into the CAMERA systems, the CAMERA team partnered with researchers and developers from the San Diego Supercomputing Center who had been involved in the Kepler project from its beginning. Bioinformaticists on the CAMERA development team created sample Kepler workflows to provide basic functionality and guiding examples for the scientists who would now be asked to create their analyses scripts as workflows. Training sessions were planned and curricula were developed to train graduate students to create and use workflows. This is an example of the ongoing leveraging and aligning work that is necessary to sustain cyberinfrastructure. Without these types of gradual sustaining changes, the CI will become less and less valuable to the community over time.

DISCUSSION

Given that cyberinfrastructures like CAMERA are in a state of ongoing change, sustainability can be seen as matter of tending to a set of infrastructural relationships to maintain productive alignments. Only by responding to new organizational arrangements, technologies and scientific needs can an infrastructure remain relevant.

However, it is also clear that changes in the cyberinfrastructure take time. If change is too rapid, or affects too much of the relational structure of the CI, it is more likely to have negative consequences. In CAMERA, we saw changes and their consequences happening over time scales of two or more years. We also see change happening at different rates in different systems. Even while the middleware has changed significantly in this time period, the computing hardware has remained quite stable. We expect that as the infrastructure matures, the overall pace of change may slow down or speed up, but we do not expect change to stop.

In the rest of this discussion, we address some of the implications of our findings for how we think about and design for sustainability of cyberinfrastructure.

Coupling Human and Technological Infrastructure

The value of infrastructures lies in their ability to connect other systems and infrastructures. The current research reveals that in the richly embedded and interconnected sets of relationships among the human and technological components that comprise cyberinfrastructures, changes in one component lead to shifts in other connected components.

A change in one component of the infrastructure cannot be understood without understanding the way that the component is connected to others. We observed, for example, a very close coupling between technological changes and changes in human infrastructure. In CAMERA, a change in organizational participation (JCVI leaving the project, CRBS joining the project) also involved a set of changes for individuals (some staff leaving the project, others joining) and technologies (switching to different middleware applications). The perturbations in the technological components then enable other changes, for example, developing new collaborations with the Kepler open source project, or shifting more of the responsibility for application development to the scientist users. These responses propagate through the networks and webs of relationships that make up the infrastructure, as each component adjusts (or is adjusted) to changes going on around it.

The *synergizing* concept reminds us that even though these complex systems have emergent properties and far-reaching effects, they are not beyond the reach of design. In the rest of this discussion, we will consider what the example of CAMERA reveals about what it means to design for sustainable cyberinfrastructure.

Ongoing Maintenance Work

Sustainability of a cyberinfrastructure over time is a process of ongoing maintenance of infrastructural relationships among people, organizations, and technologies. This maintenance work is taking place against a constant backdrop of change. A change in one component produces a spreading set of responses as other components and systems adjust and adapt to each other. These adaptations spread through the cyberinfrastructure at different rates, with some components adapting quickly (e.g., a bug fix in a script may be complete in a few minutes) while others take years (e.g. getting community buy-in for a new data standard). Making these adaptations is the work required to sustain cyberinfrastructure.

It is important to realize that this work operates not only on the components themselves but also on the relationships between them. The kind of adaptation varies depending on the circumstances. Some relationships remain relatively stable (e.g. the relationships between the CAMERA project and the stakeholders on its advisory board), with only minor changes to keep the relationships productive. In other cases, it may make more sense to end a relationship than to try to keep it active. Frequently, new relationships are cultivated, perhaps to expand the scope of the cyberinfrastructure, or replace another relationship that is no longer active. This work takes place against an always-changing backdrop of evolving stakeholder needs, developing science, new technologies, unpredictable funding, and shifting organizational attachments. A cyberinfrastructure that does not keep pace with these changes will quickly become obsolete.

This perspective helps to reframe the tension between innovation and upkeep. While we agree with Ribes and Finholt [23] that creating something perceived as new is often seen as more valuable than maintenance, the synergizing lens reveals that the day-to-day work of maintaining the CAMERA cyberinfrastructure is not fundamentally different than the work of developing new capabilities. In fact, maintenance work could be better described as continual redevelopment work.

Designing for Flexibility

Sustaining cyberinfrastructure requires continual adaptation. Ribes and Finholt find that there is a tension in infrastructure development between maintaining flexibility in the face of pressure to have concrete and detailed plans [23]. We see this tension as a design opportunity and believe there are opportunities to make flexibility and responsiveness a design consideration for the development of cyberinfrastructure. For example, an active and evolving research area is to consider how to flexibly connect heterogeneous data sources using semantic web technologies [9].

When designing infrastructures, it is important to consider both flexibility and stability. As components become more embedded within infrastructures, the networks of relationships tend to reinforce each other and become stronger and more productive [6]. Too much flexibility in a cyberinfrastructure could lead to more instances of breakdown and a loss of robustness.

Flexibility need not be uniform across the entire infrastructure. Some components and relationships can be more flexible while others are more rigid. Similarly, flexibility can be adjusted over time. We saw the CAMERA developers using this to their advantage. For example, when CRBS joined the project, most of the technological components were kept unchanged while the human and organizational components were in flux. It was not until two years later, when the human infrastructure was more established, that the team released the new version of the software. The timing of changes was carefully managed in order to maintain the balance between flexibility and stability.

Although it is somewhat counterintuitive, we believe that standardization can be an important part of designing for flexibility. In CAMERA, we saw the adoption of genomic data and metadata standards, and the use of the Kepler workflow system to standardize the representation of data analysis scripts. In both of these cases, adopting the standard freed up resources that could be used to address other areas where flexibility was more important. The key design decisions involve which aspects to standardize and which need to remain flexible. A good standard can lead to greater efficiencies, but standardization that happens before stakeholders have reached agreement can stifle necessary innovation [14].

CONCLUSION: CI SUSTAINABILITY

In this paper we have explored how one cyberinfrastructure project, CAMERA, reacted to and managed change in pursuit of sustainability. By conducting investigations at two time periods we were able to highlight aspects of change that might not be as salient in shorter time periods. We recognize that the three-year time period covered in the study is still short compared to other work that has focused on longer time spans (e.g. [23]). Certain sustainability issues are beyond the scope of this study, like how to ensure long-term funding on decade-long scales or what to do when it is decided that an infrastructure has outlived its usefulness. However, we argue that with CAMERA, we were able to see many of the same kinds of changes that cyberinfrastructures will face in the longer term: refactoring of organizational and human resources, changing demands from user communities, and the introduction of new technologies. We hope that these findings can inform not only the work of developing cyberinfrastructure, but also policy and planning discussions.

Viewing sustainability through the lens of synergizing highlights that the work involved in maintaining a productive CI involves managing the infrastructural relationships among organizations, individuals and technologies. The human and technological infrastructures are tightly coupled, and changes in one area often require adjustments in others. As CAMERA faced fluctuations in its human infrastructure, technologies, and scientific user community, the developers actively moderated the pace of change to prevent major disruptions, and paid careful attention to the balance of flexibility and stability within the cyberinfrastructure.

This case provides a better understanding of the work necessary to sustain cyberinfrastructure. Preserving the components of the CI is not enough. Myers and McGrath [20] argue that when faced with the question of what should be sustained, “sustaining capabilities and architecting to enable change are often the better choice, particularly when one considers the costs and consequences of maintaining a specific software product.” In other words, CI sustainability is less about maintaining any particular technology than it is about being prepared to accommodate technological, scientific and organizational change. The capabilities of cyberinfrastructure are a product of a relational development process that we call synergizing. Preserving the artifacts and technological elements of cyberinfrastructures is important, but much of the value of the cyberinfrastructure is lost if these are taken out of their relational context. The usefulness of cyberinfrastructure arises out of the rich interconnectedness of the human and technological infrastructures. Sustaining a cyberinfrastructure requires maintaining these relationships.

ACKNOWLEDGMENTS

This research was supported in part by NSF awards IIS-0712994 and OCI-0838601. Thanks to our participants and to our anonymous reviewers for their feedback.

REFERENCES

1. Allen, E. E., & Banfield, J. F. Community genomics in microbial ecology and evolution. *Nature Reviews Microbiology*, 3, 6 (2005), 489-498.
2. Aranova, E., Baker, K. S., & Oreskes, N. Big Science and Big Data in biology: From the International Geophysical Year through the International Biological Program to the Long Term Ecological Research (LTER) Network, 1957-Present. *Historical Studies in the Natural Sciences*, 40, 2 (2010), 183-224.
3. Atkins, D. E., Droegemeier, K. K., Feldman, S. I., et al. 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.
4. Atlas.ti v. 6. (2011). [computer software]. Berlin: ATLAS.ti Scientific Software Development GmbH.
5. Bietz, M. J., & Lee, C. P. Collaboration in metagenomics: Sequence databases and the organization

of scientific work. In *Proc. ECSCW 2009*, Springer-Verlag (2009), 243-262.

6. Bietz, M. J., Lee, C. P., & Baumer, E. P. S. Synergizing in cyberinfrastructure development. *Computer Supported Cooperative Work*, 19, 3-4 (2010), 245-281.
7. Corbin, J., & Strauss, A. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (3rd ed.). Sage, Thousand Oaks, CA, 2008.
8. Field, D., Garrity, G., Gray, T., et al. The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology*, 26, 5 (2008), 541-547.
9. Futrelle, J., Gaynor, J., Plutchak, J., et al. Semantic middleware for e-science knowledge spaces. In *Proc. MGC 2009*, ACM Press (2009).
10. Galison, P. *Big Science: The Growth of Large-Scale Research*. Stanford University Press, Stanford, CA, 1992.
11. Gerson, E. M. Reach, bracket, and the limits of rationalized coordination: Some challenges for CSCW. In *Resources, Co-Evolution and Artifacts: Theory in CSCW*. M. S. Ackerman & C. A. Halverson & T. Erickson & W. A. Kellogg, Eds. Springer, London, 2008, 193-220.
12. Glaser, B. G., & Strauss, A. L. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine de Gruyter, New York, 1967.
13. Karasti, H., & Baker, K. S. Infrastructuring for the long-term: Ecological information management. In *Proc. HICSS 2004*, IEEE Computer Society (2004), 10020c.
14. King, J. L., & Frost, R. L. Managing distance over time: The evolution of technologies of dis/ambiguation. In *Distributed Work*. P. J. Hinds & S. Kiesler, Eds. MIT Press, Cambridge, MA, 2002, 3-26.
15. Lee, C. P., Dourish, P., & Mark, G. The human infrastructure of cyberinfrastructure. In *Proc. CSCW 2006*, ACM Press (2006), 483 - 492.
16. Lee, J. W., Zhang, J., Zimmerman, A. S., et al. DataNet: An emerging cyberinfrastructure for sharing, reusing and preserving digital data for scientific discovery and learning. *AIChE Journal*, 55, 11 (2009), 2757-2764.
17. Ludäscher, B., Altintas, I., Berkley, C., et al. Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice and Experience*, 18, 10 (2006), 1039-1065.
18. Lynch, C. Big data: How do your data grow? *Nature*, 455, 7209 (2008), 28-29.
19. Mackie, C. J. Cyberinfrastructure, institutions and sustainability. *First Monday*, 12, 6 (2007).
20. Myers, J. D., & McGrath, R. E. *Sustaining capabilities not codes by architecting for innovation. In Cyberinfrastructure Software Sustainability and Reusability: Report from an NSF-funded workshop*. C. A. Stewart & G. T. Almes & B. C. Wheeler, Eds. Indiana University, Bloomington, IN, 2010, 100-102.
21. 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.
22. The Oxford English Dictionary. (1989). *OED Online* (2nd ed.). Oxford, UK: Oxford University Press.
23. Ribes, D., & Finholt, T. A. The Long Now of Technology Infrastructure: Articulating Tensions in Development. *Journal of the Association for Information Systems*, 10 (2009), 375-398.
24. Ribes, D., & Lee, C. Sociotechnical Studies of Cyberinfrastructure and e-Research: Current Themes and Future Trajectories. *Computer Supported Cooperative Work (CSCW)*, 19, 3 (2010), 231-244.
25. Schmidt, K., & Simone, C. Coordination mechanisms: Towards a conceptual foundation of CSCW systems design. *Computer Supported Cooperative Work (CSCW)*, 5, 2 (1996), 155-200.
26. Schopf, J. M. Sustainability and the Office of CyberInfrastructure. In *Proc. Eighth IEEE International Symposium on Network Computing and Applications, 2009*., IEEE Computer Society (2009), 1-3.
27. Star, S. L. The ethnography of infrastructure. *American Behavioral Scientist*, 43, 3 (1999), 377-391.
28. Star, S. L., & Bowker, G. C. How to infrastructure. In *The Handbook of New Media*. L. A. Lievrouw & S. Livingstone, Eds. SAGE Publications, London, 2002, 151-162.
29. Star, S. L., & Ruhleder, K. Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, 7, 1 (1996), 111-134.
30. Stewart, C. A., Almes, G. T., & Wheeler, B. C. *Cyberinfrastructure Software Sustainability and Reusability: Report from an NSF-funded workshop*. Indiana University, Bloomington, IN, 2010.
31. Turnbaugh, P. J., Backhed, F., Fulton, L., et al. Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host Microbe*, 3, 4 (2008), 213-223.
32. Turnbaugh, P. J., Ley, R. E., Hamady, M., et al. The Human Microbiome Project. *Nature*, 449, 7164 (2007), 804-810.