Computer Supported Cooperative Work (2010) 19:245–281
DOI 10.1007/s10606-010-9114-y

Synergizing in Cyberinfrastructure Development

Matthew J. Bietz[1], Eric P. S. Baumer[2] & Charlotte P. Lee[1]
[1]*Department of Human Centered Design & Engineering, University of Washington, Campus Box 352315, Seattle, WA 98195, USA (Phone: +1-206-543-2567; E-mail: mbietz@u.washington.edu);*
[2]*Department of Informatics, University of California, Irvine, USA*

Abstract. This paper investigates the work of creating infrastructure, using as a case study the development of cyberinfrastructure for metagenomics research. Specifically, the analysis focuses on the role of embeddedness in infrastructure development. We expand on the notion of human infrastructure to develop the concepts of *synergizing*, *leveraging,* and *aligning,* which denote the active processes of creating and managing relationships among people, organizations, and technologies in the creation of cyberinfrastructure. This conceptual lens highlights how embeddedness is not only an important result of infrastructure development, but is also a precursor that can act as both a constraint and a resource for development activities.

Key words: Cyberinfrastructure, Synergizing, Leveraging, Aligning, Infrastructure, Metagenomics

1. Introduction

Infrastructures pose numerous theoretical and practical challenges, both in studying them (Star 1999) and in designing them (Ribes & Finholt 2007). Infrastructures are composed of multitudes of heterogeneous entities and relationships; they emerge and evolve over long time periods and across great physical distances; they often simultaneously have embedded in them, and are embedded in, other infrastructures; they are the result of interactions among many and varied individuals, organizations, and other entities. This paper analyzes the roles of, and relationships between, embeddedness and purposeful action in the development of infrastructure, by studying the development of one particular cyberinfrastructure. We introduce the concept of *synergizing* to highlight the work that developers of infrastructure do to build and maintain productive relationships among people, organizations, and technologies.

Cyberinfrastructure is a specific class of infrastructure that brings together people, information, and technologies to support research. Cyberinfrastructures employ or develop cutting-edge information technologies to enable large research endeavors with potentially far-reaching impacts, endeavors that could not be undertaken without the existence of such infrastructures. Early advances in this area focused almost exclusively on scientific research, e.g., unifying functional brain imaging scans from multiple distributed work sites (Lee, et al. 2006). While

physical and biological scientists have long been avid consumers and developers of computational technologies, the scope of cyberinfrastructure has broadened to include other domains including humanities and the arts, facilitating the development of such emerging fields as digital humanities research (Davidson and Goldberg 2004). Cyberinfrastructure represents a shift toward collaborative research in distributed, technologically-supported environments, and it provides a unique research site from which to gain insights about the process of infrastructure development or "infrastructuring" (Karasti & Baker 2004).

Infrastructures represent complex sets of relationships embedded in and constrained by other systems, making it impossible to predict perfectly in advance what the infrastructure will be or how it will be used. Star pointed out that the properties of infrastructure emerge over time and through use:

> *Because infrastructure is big, layered, and complex, and because it means different things locally, it is never changed from above. Changes take time and negotiation, and adjustment with other aspects of the systems are involved. Nobody is really in charge of infrastructure. (Star 1999, p. 382)*

It is important, however, not to take this to mean that infrastructures emerge at random or are completely unpredictable. Such a stance glosses over the role of intention in the development of cyberinfrastructures and can render the day-to-day work that people do to create infrastructure invisible. Cyberinfrastructures are developed for a reason. Scientists want specific features and need to answer specific research questions. Advisory boards dictate development plans and goals, and the software engineers and other developers work to meet those goals. Cyberinfrastructures that fail to live up to expectations are unlikely to maintain their funding streams. The emergent properties of cyberinfrastructure result in part from a great deal of purposeful action. In this paper we aim to better understand the work of developing infrastructure.

As exemplified in the Star quotation above, the embeddedness of infrastructure can be seen as constraining ongoing development. Infrastructures have relationships with and dependencies on various systems, resources, and other infrastructures, which limit the autonomy of the systems and their developers. In this paper, we recognize this aspect of infrastructure development, but we also understand embeddedness as a resource that cyberinfrastructure developers can use to accomplish work. Developers draw on existing arrangements of relationships to build new infrastructure.

We explore these ideas within an ethnographic study of the work of the Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA), a large-scale, multi-year cyberinfrastructure being developed for the nascent marine metagenomics community. While the findings here arise from and relate to cyberinfrastructure, the theoretical contributions in this paper could beneficially be applied to studies of infrastructure more generally. Our investigation reveals a set of collaborative

strategies that we call synergizing, which serves as a conceptual lens that draws attention to how infrastructural work is accomplished, specifically, the purposeful creation, management, and enactment of infrastructural relationships, and the pivotal role of embeddedness as an infrastructural resource.

2. Background: cyberinfrastructure

Scientists and policy makers recognized early the potential of networked computing for scientific practice, and envisioned the collaboratory (a portmanteau of collaboration and laboratory) as a "center without walls" enabled by high-speed computer networks in support of distributed science (Wulf 1993). A significant body of research and development work grew up around collaboratories, e-Science, and virtual laboratories (Finholt 2002; Hey & Trefethen 2003; Olson et al. 2008). More recently, the term *cyberinfrastructure* has grown in prominence after the report of a 2003 NSF Blue-Ribbon Panel (Atkins et al. 2003; also, Hey and Trefethen 2005). This shift in language suggests a deliberate comparison to transportation networks or electricity distribution grids. Like these traditional infrastructures, cyberinfrastructures are complex technological and social systems (Edwards et al. 2007).

The growth of large-scale cyberinfrastructure projects reflects a trend toward more complex configurations of scientific collaboration (Sonnenwald 2007) including: 1) a movement towards large scale enterprises such as those for physicists (Galison 1997; Traweek 1988), and 2) the rise of interdisciplinarity which is related to the growth of big science but is also strongly associated with changes in funding for science and the sites and contexts of knowledge production (Gibbons et al. 1994). The gap between disciplines in modern day science is perceived as a natural place for new disciplines to evolve: "The real-world research problems that scientists address rarely arise within orderly disciplinary categories, and neither do their solutions. Thus, the information needed to solve complex research problems is distributed across disciplines and takes many different forms, physically and intellectually" (Palmer 2001). Cyberinfrastructure development is thought of as requiring interdisciplinary collaboration, particularly between technologists and domain scientists (e.g. physicists, hydrologists, biologists, etc.). In certain cases, such as the one described in this study, the cyberinfrastructure is also meant to stimulate a new discipline entirely. The interdisciplinarity and novelty of the science itself creates a challenge for designers of these systems, namely, figuring out who will be using the system and for what.

2.1. Cyberinfrastructure and infrastructuring

While infrastructures are often discussed as a thing or artifact, this conception of infrastructure is problematic. Star and Ruhleder (1996) point out that rather than asking "What is an infrastructure?", a more appropriate question is "When is an

infrastructure?". "Infrastructure is a fundamentally relational concept. It becomes infrastructure in relation to organized practices" (p. 113). In the title of their chapter, "How to Infrastructure," Star and Bowker (2002) treat infrastructure as a verb as much as a noun. This theme has been echoed by others. Karasti and Baker (2004) refer to "infrastructuring" as the ongoing process of creating infrastructure. Pipek and Wulf (2009) similarly use "infrastructuring" to "subsume all activities that contribute to a successful establishment of usages" of infrastructures (p. 450). This paper takes a similar approach: we are primarily concerned with the process and practices of cyberinfrastructure creation and use.

As many cyberinfrastructures are still developmental efforts (Lee et al. 2006), few have reached their infrastructural goals. The resulting effect is that projects tend to be too large to study completely yet small enough to entice those engaged in studies of work to attempt a comprehensive treatment. Cyberinfrastructures require ongoing development and maintenance (Edwards et al. 2007); research about how they function and whom they are serving must by necessity grapple with their emergent and shifting qualities. A development timeline of 10, 15, even 20 years presents special challenges to the CSCW and HCI communities in terms of both framing research questions and defining what it means to gather requirements. The scientists and indeed the science itself are expected and encouraged to change over the course of the development lifecycle of the infrastructure and beyond. This area of research, while full of possibility, has lent itself to two immediate approaches: 1) choose a somewhat bounded subgroup (e.g., that share a very particular and already defined set of research or development tasks) for whom to research and design (e.g., Poon et al. 2008), or 2) grapple with messy assemblages of human infrastructure and socio-technical systems (e.g., Latour 1993; Lee et al. 2006; Ribes & Finholt 2007) to identify patterns of work and cooperation that enable the whole infrastructure to emerge and function so that these patterns can be supported. These "messy assemblages" are more visible during this early phase of development, when arguments and negotiations are still ongoing and the cyberinfrastructure has yet to achieve a more stable state. This paper is a contribution to the latter category.

Cyberinfrastructure represents a new front in both science and technology. Project members are establishing advanced tools and new practices, while at the same time they are establishing new scientific disciplines. Rather than approaching cyberinfrastructure as revolutionary, we wish to understand cyberinfrastructure as it appears from the perspective of those who are creating it and to investigate the building of cyberinfrastructure as a process that entails the incremental alignment and realignment of people, processes, and tools.

2.2. Developing cyberinfrastructure

The notion of human infrastructure has done the necessary work of drawing attention to the diverse, and often invisible, collaborative structures needed to

create and support infrastructure building (Lee et al. 2006). The term human infrastructure was originally coined by Berman (2001) in a feature article to refer to:

> *A synergistic collaboration of hundreds of researchers, programmers, software developers, tool builders, and others who understand the difficulties of developing applications and software for a complex, distributed, and dynamic environment. These people are able to work together to develop the software infrastructure, tools, and applications of the cyberinfrastructure. They provide the critical human network required to prototype, integrate, harden, and nurture ideas from concept to maturity.*

Human infrastructure has subsequently been taken up and theorized by scholars to explore the variety of forms that collaboration may take in the development of large-scale collaborations such as cyberinfrastructure development (Lee et al. 2006). Human infrastructure posits that complex infrastructures come about through complex interactions among networks, place-based organizations, groups, and consortia. Human infrastructure also posits that participation takes many forms and that no single organizational form such as teams, networks, or organizations can account for the whole. Human infrastructure is complex and heterogeneous and *participation may necessarily take some or all of these forms simultaneously.* As Latour (1987) has noted before us, an approach focusing on a particular unit of analysis, and particularly to focus on one unit at the sacrifice of a lesser or greater one, is made for the convenience of the analyst and not because that is a realistic model of how science and engineering unfolds. This work, then, does not focus on a single, restricted unit of analysis, but rather seeks a more holistic understanding of cyberinfrastructure development.

Research on "tensions across the scales" depicts the landscape of activities and concerns that are prevalent in the realm of cyberinfrastructure. Ribes and Finholt (2007) describe three scales of action for infrastructure development: enacting technology, organizing work, and institutionalizing. They also identify three persistent development concerns: motivating contribution, aligning end-goals, and designing for use. Nine tensions arise at the intersections of the scales and the concerns. For example, at the intersection of enacting technology and designing for use lies the tension, "Today's requirements vs. tomorrow's users." Additional tensions include "project vs. facility," "individual vs. community," "research vs. development," etc. These tensions create a useful framework for understanding the rocky terrain that developers must navigate when creating cyberinfrastructure.

One of the tensions identified by Ribes and Finholt that is a key issue for the current paper is "planned vs. emergent." Edwards et al. (2007) describe this as the challenge of "navigating processes of planned vs. emergent change in complex and multiply-determined systems." Infrastructures sit within an arrangement of constantly evolving relationships to other systems and infrastructures. This arrangement of relationships constrains the developer. The locally optimal

solution that most closely conforms to the planned course of action may create untenable incompatibilities that isolate the developing infrastructure from other systems and infrastructures. While Ribes and Finholt, and Edwards et al. identify and describe this and other tensions, little is yet know about how the developers of cyberinfrastructures are working to meet these challenges.

2.3. Embeddedness

A key characteristic of infrastructure is *embeddedness*: it "is 'sunk' into, inside of, other structures, social arrangements, and technologies" (Star and Ruhleder 1996, p. 113). Infrastructures are characterized as much by their relationships to other systems and infrastructures as they are by any particular set of technologies. Edwards, et al. (2007) sees this as a property that distinguishes systems from infrastructures. While systems are locally controlled, infrastructures operate over a range that runs from "networks (linked systems, with control partially or wholly distributed among the nodes) to webs (networks of networks based primarily on coordination rather than control)" (p. 12). To say that an infrastructure is embedded means that the infrastructure is situated within a network, web, or other arrangement of relationships to other systems. Those relationships both enable the infrastructure to provide useful services at the same time that they constrain the modalities by which those services are provided.

An example from the electricity infrastructure may be illuminating. A large number of social and technical relationships are invoked in the design of a wall outlet. Manufacturers of wall outlets must use the same standard plug shapes and sizes as manufacturers of home appliances. The plug must also be in line with local building codes and must be able to handle the power provided from the electric grid. The existence of these relationships makes it possible for the outlet manufacturer to produce a physical object that will work within the electrical infrastructure. But at the same time, these relationships also prevent the manufacturer from making significant changes to the shape of their outlet.

The use of the concept of embeddedness here shares much with the way the term is used in economic sociology (Granovetter 1985). Like Granovetter, we are interested in the way that relational structures enable and constrain certain kinds of action. However, Granovetter is specifically focused on social networks and their impact on economic activity, whereas the relationships we explore are not limited to social networks. In other words, we are interested in both network and non-network structures (like webs or hierarchies), and we are interested in relationships among different types of entities (individuals, organizations, technologies, etc.). This is not to say that we are arguing for a generalized symmetry that does not recognize a distinction between humans and non-humans (cf. Callon 1986; Gad & Bruun Jensen 2010; Latour 1987). Our work here remains focused on the process of purposeful human action in cyberinfrastructure development.

Infrastructures are often embedded in multiple overlapping relational structures (Star and Bowker 2002). An infrastructure may be simultaneously embedded in technological networks, interpersonal networks, organizational networks, etc. In the electricity example above, there are relationships of technologies (voltages and plug shapes), legal relationships embodied in building codes, organizational-level relationships when manufacturers and utilities participated in the development of standards, etc. This multi-faceted embedding will play a particularly important role in this paper—we will see that developers sometimes employ a tactic of using relationships from one organizing structure (e.g. groups, networks, organizations, etc.) to create new connections to other structures.

Embeddedness can be both a limitation and a resource. On the one hand, the embeddedness of infrastructure constrains action. In the electricity example above, if plug manufacturers created an innovative plug that did not work with standard electrical sockets, it would have little chance of success. Infrastructures do change over time, but "the installed base of a particular infrastructure carries huge inertia" (Star and Bowker 2002, p. 158). At the same time, much of the value of infrastructures lies in the relationships they embody. Infrastructures can benefit from "network externalities," those situations in which the value of a good or service increases as more people use the same good or service (Katz and Shapiro 1985). This concept is often described with examples such as telephones and fax machines, which become more valuable as more people own them. We can see the same dynamics in cyberinfrastructure resources like large genetic databases or standards for data interchange (Star and Bowker 2002) whose value depends on wide-spread use.

Embeddedness is a valuable result of the development of infrastructure, but it is also a resource for that same development process. Developers of infrastructure can draw on these arrangements of relationships to build, maintain, or strengthen other relationships. The simplicity with which we have described our electricity example belies the difficulty of creating and maintaining embedded infrastructures. Each system is coupled to numerous other systems. The landscape of connections is dynamic, and often contains unresolved conflicts. Managing these alignments between different organizing structures is a key activity in building infrastructure, and requires a great deal of innovation and hard work. Embeddedness can be an important resource for doing this work. We refer to this active, strategic work of managing multiple relationships for infrastructure development as "synergizing."

3. Synergizing

In order to better understand how cyberinfrastructures become embedded in other systems and infrastructures, we draw on the concept of *synergy*—increased effectiveness produced as a result of combined action or co-operation ("The Oxford English Dictionary" 1989). We find that synergy is part of the common

lexicon within the world of cyberinfrastructure, and was mentioned by many of our research participants. While the synergies themselves are important, we are more interested with the process of creating and maintaining productive socio-technical relationships, which we call *synergizing*. The concepts of *synergy* and *synergizing* are frequently dismissed as fancy, superficial buzzwords for working as a team or mere reaction to a reduction of resources. Instead, our research posits the concept of synergizing as a particular class of collaborative strategies undertaken in the milieu of infrastructure building projects that are large, distributed, loosely formed, and long-term.

Our goal in this paper is to capture the intentional day-to-day activities that accumulate into infrastructural embeddedness. *Synergizing,* along with its component subprocesses of *aligning* and *leveraging* (defined below), serve as a useful analytic lens to help expose the work required to create cyberinfrastructures. Our focus is not on characterizing any particular synergy, but rather on understanding how these interactions come into being, are maintained, and can be made productive. It should be noted that when we use the term "developer" in relation to cyberinfrastructure, we are using a broad definition that is not limited to someone who builds hardware or writes software. Instead, we use the term to refer to anyone who is intentionally doing work that creates cyberinfrastructure. Thus, developer can refer as much to a program manager at a funding agency or a scientist who contributes data as to a computer scientist or software engineer.

We approach synergizing as a broad concept that includes strategic collaborative undertakings in pursuit of greater combined effects than individuals, groups, or organizations could effect on their own. Synergy can arise from bringing two groups together in a collaborative relationship, or it can come from the linking together of two pieces of software to produce a more capable system. Synergy arises from bringing together already-existing entities, rather than "from scratch" development of new entities or growth of a single entity. However, something new may be created as a result of the synergy.

The process of developing infrastructure requires building relationships among different entities. We must consider both a diverse set of entities and a diverse set of relationships. Here, the entities we are mostly concerned with include computational and scientific technologies, people, organizations, and communities, although depending on the context, we could also invoke other entities like teams, governments, etc. The relationships among entities also vary depending on the context. Two individuals may have worked together in the past, or they could have a personal friendship. An organization might sub-contract to another organization. A technology may use another as a component, or draw on services provided by another. These relationships can also exist among unlike entities. An organization may hire a person, an individual may have expertise in a particular technology, or an organization may use a particular technology to manage its business practices. We are not overly concerned with creating comprehensive lists of relationships and entities. Rather, we aim to understand how relationships

between entities, including but not limited to those in human infrastructure, are made productive and used in the development of cyberinfrastructure.

Nardi et al. (2002) develops the concept of "netWORK" to describe the work required to create, maintain and activate personal social networks. The authors describe a new class of netWORKers whose work is not organized around traditional teams or groups, but rather gets done through personal social networks. The current study takes inspiration from Nardi et al.'s focus on the work required to build, maintain, and use a network. But we find that in order to understand the development of cyberinfrastructure, it is not enough to consider only personal networks. Cyberinfrastructures are embedded in complex socio-technical structures, and we must consider the relationships among the various constituent social and technical entities.

Synergizing has much in common with, but is distinct from, articulation work. The primary concerns of synergizing are ensuring that a common field of work exists and ensuring that work can be done at all, as opposed to ensuring that work goes well and that complexity is controlled within an existing field of work. In the Discussion section below, we address in more detail the relationship between synergizing and articulation work.

We define two key subprocesses of synergizing: *aligning* and *leveraging*. Aligning is the work that developers do to enact a relationship in a way that enables it to produce, and to function within, the nascent cyberinfrastructure. In the context of articulation work, Strauss (1988), drawing on Blumer (1969), defines *interactional alignment* as "the process by which workers fit together their respective work-related actions." We build on this definition, but expand it to include not just the fit between workers, but the fit or compatibility between any type of entities. These compatibilities take many forms. Collaborators need to develop shared understandings of key concepts (Spencer et al. 2008), and manage trust and conflict in their collaborations (Jones and George 1998). Public universities and private corporations may have different policies about making data and findings public, but successful collaborations can be built when agreements (often involving lawyers and contracts) can be reached about ownership and publication of results. A software engineer who wants to use two component technologies in the same system may align them by creating an application programming interface (API) that allows them to interoperate. Fit can also be an issue between entities of different types. In order for a hospital to use a particular database system, significant work has to be done to make sure that the technology is compatible with organizational and governmental policies concerning the confidentiality of patient data. It is important to note that alignment need not be perfect; our focus is on the work necessary to create enough compatibility between entities so that the relationship can be productive.

Leveraging, the second key subprocess of synergizing, is of particular importance for work creating infrastructure. Leveraging is using an existing relationship with a person, artifact, or organization to build or strengthen a

productive relationship with another person, artifact, or organization. It provides the analytic link between the management of individual relationships and infrastructural embeddedness, explaining how embeddedness can be seen both as a result of and a resource for development work. Here we use the lever—a simple machine that involves a rigid object with a fulcrum, much like a seesaw on a playground—as a metaphor for this component of synergizing. In much the same way that pushing on one end of the seesaw results in a force being applied at the other, leveraging is a way to produce indirect effects via series of relationships. In order to gain access to cutting edge technology that is not commercially available, a cyberinfrastructure project may create a partnership with a research project located at the same university that is developing advanced high-speed networking. The cyberinfrastructure project leverages its existing connections to the university to make it easier to find potential partners and develop relationships with them (see Figure 1). Or a funding agency may leverage its existing relationships with grantees to develop new collaborations (see Figure 2). Leveraging simultaneously takes advantage of embeddedness by drawing on existing relationships, and creates embeddedness by enacting new relationships.

In both of these examples, we see already-existing relationships being used to create new synergistic interactions. The interactions are synergistic because they can be made to accomplish more than the existing relationships alone would. The result is that an indirect relationship becomes a more direct one and the relational structure becomes more dense. Leveraging, like synergy, is often dismissed as a buzzword, but here, we find that it usefully captures the work of using existing relationships to aid synergizing.

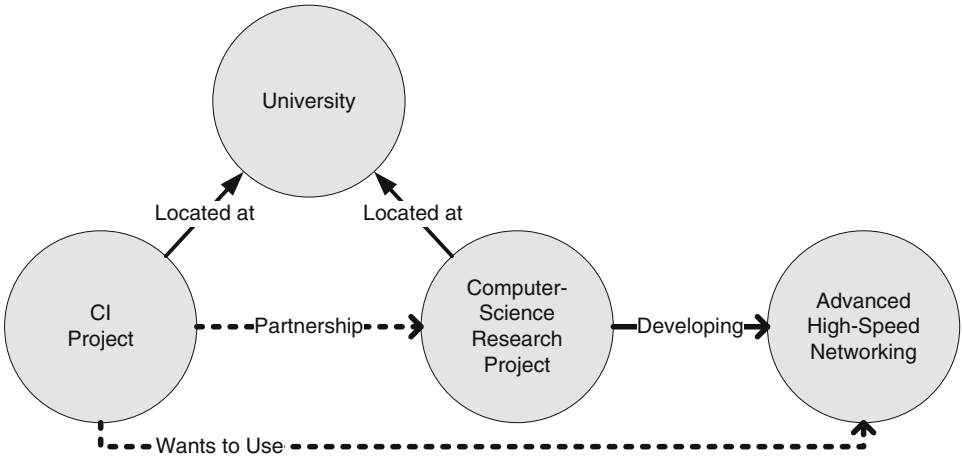

Figure 1. A cyberinfrastructure project can develop a partnership with another project at the same university to gain access to cutting-edge technology.

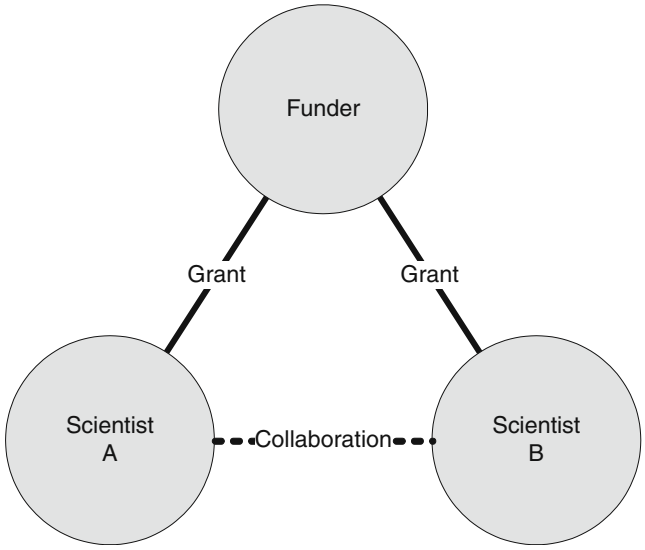

Figure 2. A funding agency uses its existing relationships with scientists in order to develop a product collaboration among scientists. Scientist A and Scientist B both have relationships with the funding agency, and these relationships can be leveraged to create a more direct relationship between the two scientists.

The rest of this paper develops these ideas further and exemplifies them using data from a particular instance of cyberinfrastructure development. We first describe our methods and the study site, and then we present two examples of synergizing activities. Finally, we discuss implications of this work for studying and developing cyberinfrastructure.

4. Our study and methods

In order to investigate the development of cyberinfrastructure, we employed qualitative research methods including ethnographic observation, interviews, and analysis of artifacts such as documents, web pages, databases, etc. This approach consisted of entering into sites involved in the production of scientific research, technological artifacts, and other resources, getting to know the people, participating in the daily routines of the settings, and observing what was going on. Our goal was to observe ordinary conditions, responses to events, and experience events ourselves as much as possible in order to understand "social life as process" (Emerson et al. 1995).

Our initial goal was to investigate cyberinfrastructure development practices through an in-depth study of a single project, the Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA). We describe the project in detail in the next section. At the time of writing, we have been studying the CAMERA development project and other metagenomics

cyberinfrastructure efforts for over two years. We were granted access to the project site by CAMERA's directors, and we were introduced at a regular project meeting. Our role was as outside observers, and we were described to project staff as academic social scientists studying the development of cyberinfrastructure. In a four-month period of on-site observation at CAMERA's headquarters in the first half of 2008, we observed seven regular group meetings of the development team, six scheduled subproject meetings, and numerous ad-hoc meetings. During this time, one of the authors had an assigned desk in the development team's work area, and spent more than 70 hours in on-site observation. Six of the full-time developers had desks in this open-plan work area, and other team members had 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 provided the opportunity to ask developers about current work, or to provide explanations and context for other events.

Over the course of the study, we conducted thirty-four in-depth interviews with twenty-seven individuals. Our initial interview sample focused on the members of the CAMERA development team, including anyone who came to weekly development meetings, anyone identified as a member of the project by the project director, and anyone listed as project staff on the CAMERA web site. Of this group, we were able to interview all but four individuals who were unavailable or did not wish to participate. We developed our participant pool through "snowball" sampling, in which we asked each interviewee to list other key individuals they worked with as part of their CAMERA activities. Guided by our ongoing data analysis, we expanded our scope to include additional categories of participants, including CAMERA's funders (staff at the Gordon and Betty Moore Foundation), microbiologists and bioinformaticists who were involved in metagenomics research, developers of other similar (and sometimes competing) systems to support metagenomics research, and members of a standards development consortium in which CAMERA participates. In addition to interviews with these participants, we also conducted over 30 hours of observation in metagenomics laboratories and attended related conferences and workshops.

Interview transcripts, field notes, and various indigenous documents were analyzed using a grounded approach in which categories, concepts, and theories were developed in a process of ongoing dialog with the empirical data (Corbin and Strauss 2008; Glaser and Strauss 1967). Analysis consisted of an iterative process of coding the data and writing memos on developing themes. We began with "open coding" in which we developed codes to capture salient activities and ideas. As coding progressed, we wrote memos to expand on the emergent phenomena and to begin to link together themes and categories (Emerson et al. 1995). New data were coded as they were generated, adding to ongoing theory development. As the corpus

grew larger, we also returned to the already-coded data to further code for newly developing categories. Through this process, we gradually developed the theoretical basis for this paper.

5. Study site: CAMERA

Our primary research site was the Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA), a large-scale, multi-year project to provide cyberinfrastructure tools and resources, and bioinformatics expertise to the *metagenomics* community. These scientists are drawing on the vast and rapidly growing corpus of metagenomic information to understand links between microorganisms' genetic composition and their environments. Metagenomics is frequently described in hopeful terms, with promises to accelerate understanding of biology and deliver novel biological solutions to important societal challenges in health care, energy, and the environment (National Research Council (U.S.) Committee on Metagenomics: Challenges and Functional Applications 2007; Seshadri et al. 2007). According to the project website, "CAMERA is making accessible raw environmental sequence data, associated metadata, pre-computed search results, and high-performance computational resources. It is based on innovative cyberinfrastructure leveraging emerging concepts in data storage, access, analysis, and synthesis not available in current gene sequence resources" (Sides 2007). Our study participants also identified the importance of engaging the community of metagenomics researchers in identifying and helping prioritize implementation of services and datasets. The CAMERA project represents a significant investment in a particular vision of the future of the biological and environmental sciences, and it is intended to serve as a model for other disciplinary sciences as they adopt cyberinfrastructure.

Metagenomics represents simultaneously a developing set of scientific practices, a particular way of understanding the relationship between organisms and their environment, a focus on populations rather than individual organisms, and a growing scientific community. An NRC report calls metagenomics "a new science" (National Research Council (U.S.) Committee on Metagenomics: Challenges and Functional Applications 2007). Metagenomics takes advantage of new DNA sequencing technologies that enable DNA to be extracted directly from communities of environmental microorganisms, thus sidestepping the need for laboratory culturing or isolation. Currently, there is little information on the vast majority of microorganisms present in Earth's different environments due to the difficulty of culturing them in the laboratory. Proponents suggest that metagenomics has tremendous potential in the development of new biocatalysts for industrial and medical applications and as a way to gauge changes in biodiversity and environmental health (National Research Council (U.S.) Committee on Metagenomics: Challenges and Functional Applications 2007).

Metagenomics, however, is far from ready-made science (Latour 1987). There is no agreement among the scientists we interviewed about the disciplinary status of metagenomics, and many are reluctant to align themselves too closely with it. One scientist, when asked if he considers himself a metagenomicist, answered:

> *No not really. I kind of hate it. I mean, I've probably done more of it than most anybody else on the planet, but, yes, I don't like it very much. It's a tool for me; very much so. So people definitely think of me that way, but no.*

In addition to the high-level issues of developing this "new science," there is significant effort being put into associated on-the-ground activities. Committees are debating over standards for metadata, new data analysis tools are being developed, textbooks are being written, and funding agencies are developing new grant categories. Because of the sheer volume of metagenomics data, new techniques and technologies must be and are being developed to generate, store, analyze, and disseminate the data. This study presents a unique opportunity to examine not only the process of creating infrastructure but also connections to the process of doing science, of science in the making.

CAMERA is a relatively new project, receiving funding only as of 2006. When we began this study, some collaborations were already occurring between institutions, within institutions, and between the CAMERA project and the intended end-users: microbial biologists, metagenomicists, ecologists, etc. As such, the project is an excellent site for understanding the development of cyberinfrastructure. The project is mature enough that its basic form has been established, but it has not yet reached a stable state where scientific and technological controversies have been closed (Pinch and Bijker 1984).

5.1. Marine metagenomics and the CAMERA technology

Before describing the specifics of our field site, it is important to provide some brief background on the scientific context. In marine metagenomic studies, samples of ocean water are filtered to extract the microorganisms. The DNA of each of these microorganisms ranges from a few thousand to a few million base pairs, but DNA sequencers can reliably read only short (fewer than 300 base pairs) segments of DNA. "Shotgun sequencing" overcomes this limitation by randomly breaking each strand of DNA into many smaller segments. Sophisticated computer programs then look for overlapping portions of these smaller segments to reassemble the larger DNA strands. One of our informants compared this task to mixing up all of the pieces of an unknown number of jigsaw puzzles of different sizes, and asking a computer to assemble all of the puzzles again. The task is made more difficult due to inaccuracies in the DNA sequencing, missing segments, segments of DNA that are widely shared across many organisms, and genetic variations within the same species.

Once the DNA has been sequenced, scientists may begin characterizing the water sample by analyzing which organisms are present, the diversity of organisms, or the functional characteristics of the genes in the sample. One of the promises of metagenomics is to be able to look across datasets to ask questions about how the characteristics of an environment shape the microbial population, and vice versa. This requires not only large databases of sequence information, but also metadata about the environment from which the samples were collected.

The CAMERA project is one attempt to meet this field's computation and data requirements, providing access to high performance computing clusters and more than 150 terabytes of data storage. CAMERA staff and their collaborators are developing specialized metadata and 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 (Smarr et al. 2003). Anyone can register and gain access to most of CAMERA's tools and data at the project website.

5.2. Infrastructuring and CAMERA

In our analysis, we approach CAMERA as a locus of infrastructural work. One is tempted to ask, "Is CAMERA infrastructure?" However, this question brings us back to a focus on the artifact rather than the activities of creating infrastructure. As discussed above, we are interested in *infrastructuring*. But even if we borrow Star and Ruhleder's (1996) question, "When is infrastructure?", the answer for the majority of usages is probably that CAMERA is not now, or at least not yet, infrastructure.

On the other hand, CAMERA development is intentionally infrastructural. The name of the project includes "cyberinfrastructure." In interviews, CAMERA's staff spoke of both creating an infrastructure and of being part of larger already-existing scientific computing infrastructures. CAMERA is intended to be a global resource that supports local metagenomics research across a wide variety of sites and research questions, that transparently links to other existing and future resources, and becomes a primary resource for anyone conducting metagenomics studies. CAMERA's developers seek for it to be embedded—technologically, socially, institutionally—in the field of metagenomics. While the jury may be out on whether CAMERA is ultimately successful in its infrastructural goals, it is a fertile site for understanding infrastructuring.

In the next section, we describe the findings from our data, using quotations from the interviews as illustration.

6. Synergizing activities in CAMERA

This section presents an account of various synergizing activities in the development of the CAMERA cyberinfrastructure. These examples are compiled from a combination of participant reflection and our observations of the

development process. A key finding that emerges from our analysis of interviews and field notes is that cyberinfrastructure results from intentionally leveraging existing relationships and creating alignment among interacting entities. Developers of infrastructure are adept at managing relationships among diverse entities including people, organizations, communities and technologies. Synergizing operates within multiple organizing structures, and it both results in and is influenced by infrastructural embeddedness.

The examples presented here cover two areas of infrastructuring activity. The first example presents a retrospective account of the early conception and funding of the CAMERA project. Rather than limiting our concept of developers to those people who directly interact with technological artifacts, we focus on any work that develops, or is intended to develop, infrastructure. This approach allows us to see that, although a program manager at a funding agency may be working on different arrangements of relationships than a database administrator, the work of both involves significant leveraging and alignment.

The second example looks at what might be considered a more typical system development activity: the design of database schema. We begin with the work of a single database programmer who is tasked with developing the database, but it becomes clear that in order to develop this piece of infrastructure requires the creation and management of multiple social, organizational, and technological relationships. In both of these examples, embeddedness is simultaneously an input into and a result of synergizing.

6.1. CAMERA's founding

While the founding of CAMERA could be glossed superficially as a non-profit funding agency creating a project to provide a service, our research demonstrates how the project's inception involved extensive synergizing work at a number of different levels. CAMERA is, from its very beginnings, intentionally embedded in a web of relationships with other infrastructures and systems. In this example we highlight how aligning and leveraging processes are crucial to building an infrastructural network of technologies, individuals, communities, and organizations.

The Gordon and Betty Moore Foundation (GBMF) is a private foundation based in the San Francisco Bay Area that focuses on environmental conservation and scientific research. The foundation has an endowment of $6 billion, and in 2007 awarded $230 million in grants (http://moore.org/faqs.aspx). GBMF organizes most of its funding around initiatives, which are designed to make a "transformative" impact in a scientific field through coordinated programs of funding. By 2003, program managers at GBMF felt that advances in microbial analysis techniques and publication of the first marine metagenomic datasets held great promise for changing marine science and our understanding of the ocean. That year the GBMF founded its Marine Microbiology Initiative (MMI) and

began funding microbiologists to work in this area. But while the laboratory methods were rapidly advancing, their awardees were reporting that requisite computational technologies were not yet in place. A project member described the birth of the CAMERA project this way:

> *The community of principal investigators basically said, "Look, there's all these [metagenomic] data coming down…. The existing databases are simply not capable of providing us with the ability to do what we need to do with these data. You've got to do something about this. Because otherwise all of these data will be lost to us or to the scientific community because the ability to query on these data will just be gone. It won't happen if you don't do something."*

The GBMF decided to address this need by funding a cyberinfrastructure to support both their own scientists and the wider microbiology community. Unlike many government grant agencies, the GBMF does not accept unsolicited proposals. Instead, GBMF identifies projects that it wants to fund and determines which people and organizations it wants to complete those projects. Representatives of the foundation approach potential grantees directly and work with them to develop the project proposal. One recipient of GBMF funding described their process this way:

> *So it's, you know, like three men in dark suits turn up and knock at your door with a briefcase. It's like we've got a deal for you. They decide what specific scientific endeavors they want to engage in…. You know, they figured out they wanted to be involved in marine biology or marine microbiology and marine ecology. But then within that realm they decide which particular initiatives they will engage in on any given year with a budget. And they do that based on advice from both their Advisory Boards, their Boards of Governors and as well as their sort of collected range of existing funded principal investigators. So, they sort of gather all of this information and recommendations and then they act on it. And as I said, they decide what needs to be done and by whom.*

GBMF approached the California Institute for Telecommunications and Information Technology (Calit2), a state-funded research institute at the University of California, San Diego (UCSD) to develop the CAMERA project. A GBMF administrator told us that choosing Calit2 for this project created the opportunity to leverage already-existing technologies, networks, and expertise:

> *[Calit2] had the cyberinfrastructure, at least knew how to build it. They also had access to people who understood the science. And so it seemed like a logical place to go, you know, one-stop shopping, if you will.*

Calit2 had a history of working on cyberinfrastructure projects and dealing with large amounts of data. They had experience with building robust high-performance computing infrastructures and access to a wide variety of already-

developed technologies. While Calit2 had significant experience with cyberinfrastructure, it did not have much experience with metagenomics and microbiology. As part of the GBMF grant, the project built a formal working relationship with the J. Craig Venter Institute (JCVI). JCVI is a leading research institute in genomic sciences, and pioneered data collection and analysis methods in metagenomics. Beyond their expertise, JCVI could also bring a unique resource to the project. JCVI produced one of the first major marine metagenomic datasets from the Global Ocean Sampling (GOS) expedition. To collect this data, a ship traveled around the world for two years, taking samples of microorganisms at regular intervals. JCVI then analyzed the DNA from these samples (with partial funding from GBMF), producing a 7.7 million sequence dataset. Another GBMF administrator described the motivation for building a collaboration between CAMERA and JCVI this way:

> *So I think part of it was, of course, the GOS was ongoing. So the data was in Venter's hands. But also, the JCVI was an enormously important DNA sequencing center during the era of the human genome sequencing, right? So they had a lot of horsepower there to build that specific DNA databases stuff that UCSD didn't have. UCSD—those guys like [senior scientist at Calit2] hadn't been working with DNA sequence information all his life. He'd been working in hardware and supercomputing stuff…. but he didn't have any, as far as I am aware, was not well versed in DNA sequence database applications. And that's where the J. Craig Venter gang came in very handy. Because not only did they generate the sequence data, but they also knew how to build databases to harbor the data, analyze it, and—because of all of their experience with genome sequencing—oh, they had the people. They had the knowledge. Basically, they had all the right stuff, including hardware and human capital, and data, of course.*

From the funder's point of view, the collaboration between Calit2 and JCVI was seen to be complementary: Calit2 brought a great deal of technical knowledge of high-performance computing and cyberinfrastructure, while JCVI could provide both the metagenomic data and expertise in managing and analyzing these datasets. Together they would be able to build something that neither would be able to build on their own.

This example begins to show that GBMF, through its representatives, approached cyberinfrastructure development as an exercise in building configurations of social and technical relationships. The foundation creates financial and contractual relationships with and among its grantees. The project was designed around a collaborative relationship that linked the microbiology and metagenomics expertise and data from JCVI with the high-performance computing resources and expertise at Calit2. However, in order to create these new relationships (among GBMF, Calit2, and JCVI), the participants drew on already-existing relationships in multiple, overlapping networks. For example,

an administrator at GBMF described the selection of Calit2 to build the infrastructure this way:

> *I've known the Calit2—well, I personally have an involvement with high-performance computing for quite a long time. And in fact, I first met [a senior investigator at Calit2] when he was a young professor…. And then over a period of time, I had a close contact with several of the super computer centers, mainly with the one in San Diego…. And so I got to—I was kind of part of that high performance computing community and then… I was involved in the competition for the second round of the super computer centers. So we kept in contact and I knew most of the people in that high performance computing community. And then subsequently, I don't remember exactly when, when the University of California started the Science and Technology Centers, they did four of them that was a special initiative by Governor Davis at the time. I was asked to be on the review committee for those. And so the one that emerged, at least in my mind, as the strongest. I think everybody agreed at the time, it was the strongest, was Calit2. And so because of that and subsequent reviews that we did after it had been in operation, I got to know what was going on there quite well. And so it seemed a logical place to turn for the kind of effort that we had in mind for this CAMERA database.*

Here the program officer leverages personal relationships to forge new organizational relationships to facilitate the building of a new infrastructure. This program officer was able to use an existing network of relationships as a resource to help build the new organizational relationship. Figure 3 is a visual representation of the network described by the GBMF administrator in the quotation above. This network is not the purpose of or reason for creating a relationship between GBMF and Calit2; rather, it is a resource that GBMF can draw on in the creation of cyberinfrastructure.

On the surface, these relationships serve an important but straightforward information function—the grant administrator knows who might be able to build the necessary high-performance computers because of his prior experience in the field. But leveraging has a deeper purpose as well: the existence of a prior relationship implies that a certain amount of alignment work has already been done. For example, because GBMF funded JCVI for another project, the organizations had already dealt with potential problems ranging from how much overhead can be charged on the grant to where to send the checks. Calit2's prior experience with other cyberinfrastructures suggests that the project members have probably already dealt with some of the inherent tensions of balancing the needs of domain scientists with those of computer scientists (Spencer et al. 2008). Here, GBMF leverages its separate relationships with JCVI (through prior funding) and Calit2 (through personal connections) to create a new relationship between the two organizations. However, in order to make the relationship valuable and productive, the participants had to work to bring the organizations into alignment.

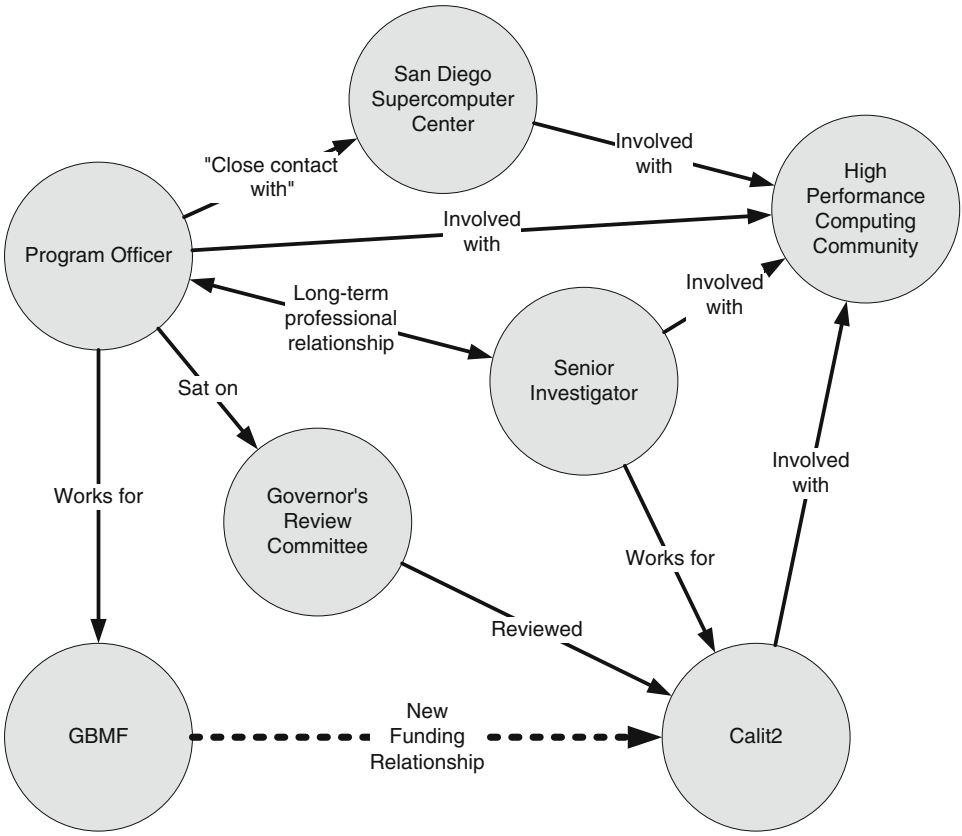

Figure 3. The network of existing relationships (*solid lines*), as described by the GBMF program officer. In the context of the CAMERA cyberinfrastructure, these relationships served as a resource that GBMF could leverage as it worked to create a relationship with Calit2. (Arrows in the diagram represent only the direction of the verbs and do not imply a directionality for leveraging. In other words, Calit2 can leverage its relationship with UCSD, but UCSD could also leverage its relationship with Calit2.)

One of our informants told us in the quotation above that GBMF "decides what needs to be done and by whom," that same informant also told us about the negotiation process that goes on:

> *I mean, essentially that came in the original negotiations for who would do what in the project. So, Moore knocks on your door and says, "Will you send us a proposal?" They don't just hand you the money…. So, fundamentally then that was a—there was a negotiation between… Calit2 personnel and JCVI personnel and somehow between them they sort of agreed upon who would do what.*

GBMF could not simply snap its fingers and create productive relationships between JCVI and Calit2. Alignment work had to be done in order to make the

new relationship between the two organizations productive. Many people were involved in negotiations about what would be done and by whom and how, both before the collaboration was formalized and as the relationship continued.

This example shows how one entity, GBMF, leveraged its organizational, financial, and personal relationships with two other entities, Calit2 and JCVI, to create a new relationship between them, as well as how that new relationship was made both valuable and productive. This synergizing work helps CAMERA become infrastructurally embedded, that is, embeddedness is produced as a valuable result of infrastructuring. The next example illustrates that these same synergizing activities form the basis of infrastructuring even in a very different arena of work, in this case, the creation of a DNA sequence database.

6.2. Building a community repository

Metagenomic methods require extensive databases of known genetic sequences. Building such a database is one of the key CAMERA development activities. CAMERA's database developers want to set up the database to best represent the structure of metagenomic data so it can be populated with data from scientists and other databases. In order to accomplish the creation of the database, the developers not only have to leverage existing socio-technical networks to create new relationships, but they also have to form various technical and social alignments to make those relationships productive.

Scientists in the genetic sciences require access to shared data from other scientists. These fields (genetics, genomics, metagenomics, etc.) have strong community norms that DNA sequence data is a public good and should be shared freely within a reasonable time after collection. Among the scientists in our study, making data public is considered routine. These social norms are backed up by commitments from journals in the field not to publish scientific results unless the DNA sequence data used in the analyses have been submitted to a publicly accessible archive (Marshall 2001). In order to support this data sharing, large sequence databases have been developed. While there are many different sequence databases in operation, GenBank is probably the largest and most famous of them and has been in operation since the early 1980 s (National Center for Biotechnology Information 2008).

Analyses of metagenomic datasets frequently begin by comparing new genetic sequences against large databases of known sequences to identify the possible species, classify the functional qualities of the metagenome, or understand the evolutionary history of organisms. Metagenomic scientists want and expect the large centralized sequence databases to contain an accurate and comprehensive collection of known DNA sequences (Bietz and Lee 2009). However, databases like GenBank were not originally intended for metagenomic data. Their data structures do not conform well to the unit of analysis in metagenomic studies, and the associated tools do not meet the needs of metagenomic scientists.

Additionally, because of the scale of metagenomic analyses (which can often take days or weeks to run on high-performance computing clusters), it is important to make sure that the database is as efficient as possible for the task at hand.

6.2.1. *Importing data*

One programmer on the CAMERA development team was given the task of creating the database schema and building mechanisms to import genetic data into the system. In order to understand the database requirements, he worked closely with a small number of datasets that would serve as test cases. In order to get access to these datasets, the CAMERA project needed to form relationships with the scientists who had produced the data. These relationships were forged by a more senior administrator on the CAMERA project, who leveraged existing networks of relationships to create relationships with the "test case" scientists. Some of these early contributors were also funded by GBMF, and their grants required them to make their data public through the CAMERA system. The administrator also spoke of the potential for these "test" relationships to endure over longer time spans or link CAMERA into other important networks. For example, they wanted to build data "pipelines" with specific sequencing centers to streamline moving data into the CAMERA database, and one of the test cases was chosen specifically because its data was generated at one of these key sequencing centers. It was hoped that developing a relationship with one scientist could provide an opportunity to build more substantial relationships with the sequencing organization in the future.

Once the database programmer had the datasets in hand, he began to figure out how to parse and import the data. At first it seemed there was a ready-made solution for importing the data. The data was all submitted in a standard file format called FASTA format (National Center for Biotechnology Information n.d.), which was a widely accepted method for sequence data interchange. Even so, the programmer found that in practice, not everyone follows the standard in the same way. For example, FASTA files contain a "header" area to provide descriptive information at the beginning of the file, but scientists use the header in different ways. The programmer told us:

> *The header files vary… some investigators are using them, others are not, you know. In one instance, we got an investigator's data and all the header line was was a number—a serialized number—relevant only to that machine, so, you know, it wasn't very meaningful. Another PI put other data within the header line…. And that's where the variant data comes from. It's a problem when you put it into the database because everyone does it differently and you want to make sure that you get meaningful data, everything you can from the header line. Ultimately it requires all these little programs to look at special cases.*

Standards support coordination by serving as reference points to which stakeholders can align themselves, but even with a standard that is as widely adopted as the FASTA format, coordination is neither perfect nor stable. The developer must engage in significant alignment work. This alignment work can involve adjusting local practice to external demands, but it frequently also involves influencing others to adjust their systems. On the one hand, the developer is writing "little programs" that translate from one data format to another, creating compatibility between the scientists' data and the CAMERA database. On the other hand, he is simultaneously giving feedback to scientists about how best to format the data, and creating templates that scientists can use to format their data in an appropriate manner. This alignment process takes place as part of synergizing: the developer is building and strengthening relationships in order to solve the problem at hand.

The FASTA format standard also provides an example of how embeddedness can act as a constraint. In working with the test cases, the CAMERA programmer found that the FASTA format is less than ideal for metagenomic data. The need for sophisticated automatic translation and significant manual data preparation could possibly be lessened by developing a file format that allowed better representation of population-level relationships and linking to environmental metadata. But the FASTA format is deeply embedded in the genetic sciences. It is used by GenBank and most other genetic sequence databases. Most (perhaps all) commercially-available DNA sequencers produce FASTA formatted data files. Many data analysis tools are built around the format. Rejecting the FASTA format would mean damaging CAMERA's relationships with scientists, sequencing centers, and other databases. While it can serve as a resource, there are many instances where embeddedness, sometimes in the form of burdensome legacy systems, can act as a hindrance.

6.2.2. *Metadata standards*

There are also cases where there is not even an imperfect standard to use in the database, as is the case with metadata. This is an area where the CAMERA infrastructure was being co-developed with a scientific practice. Scientists want to use metagenomics to answer questions like, "How does ocean temperature affect the composition of microbial communities?" In order to answer that kind of question, it is necessary not only to have sequence data, but also associated contextual data or "metadata" that provide information about the environment from which a sample was collected. But when the developer tried to develop a metadata schema for the database, he discovered that there was no agreement among scientists about what the metadata should be.

> *There's the pure [DNA]base-pair kind of stuff, and then there's the metadata, and the metadata's just a complete mess. It's all over the map. I did a chart between the [meta]data that we've gotten and if you did a Venn diagram of*

that chart, you would just have a little thin sliver of things that they all have in common. There's a big question with the database on how we do this; how we make this work and how we use this data so that it's usable. And there's a little bit of a gap between us and the scientists on, you know, what is needed and what's important. And there's even a gap between the scientists on what they consider important and usable. So a lot of this is just up in the air.

In this instance, the developer is not only facing a lack of alignment between CAMERA's technologies and scientific needs, but also a lack of agreement within the scientific community itself. At the time, a standard for metagenomic metadata was in development, but this was not something that the database programmer could simply plug into the CAMERA database. The standard would not be useful without community-wide buy-in and ongoing development as the science and technologies advance, i.e., without significant alignment work. This became another opportunity for synergizing.

Generating the kind of standard that CAMERA needs for its database requires negotiations within the metagenomics community about which research questions are most valuable, how to fund the collection of additional metadata, and how to ensure compliance with these decisions once they are made (Bietz and Lee 2009). The Genomic Standards Consortium (GSC) is an organization that serves as a venue for creating standards and agreements among members of the genomic and metagenomics communities and various cyberinfrastructures (http://gensc.org). The GSC is composed of genomics and metagenomics researchers, bioinformaticists, database developers, and other stakeholders working toward a set of standards for data and metadata. The GSC recently published its first metadata standard (Field et al. 2008), and is working not only to develop further this standard and other standards, but also to encourage their adoption and use. CAMERA decided to join the GSC to participate in the standard development process. Joining the GSC was a way for CAMERA to leverage GSC's extensive network of relationships with metagenomics scientists and cyberinfrastructures. One senior CAMERA project member described CAMERA's involvement with the GSC:

I came to realize that the Genomic Standards Consortium is extremely important for CAMERA; that it shares goals with the Moore foundation, with the U.S. federal government, and with scientists everywhere who want to have guarantees that no data is lost, that different databases can readily exchange data, that there are clear standards for people collecting whole genome data or metagenomic data. And the data—you know, the information in the Genome Project belongs to all of science and all of society…. One thing I think the GSC can help do is ensure that people working on—you know, that there aren't seven blind people looking around this piece of this elephant; that everybody's—people are looking around different parts of the elephant, but they're communicating to each other. They each can see the whole elephant, but they specialize in their pieces and they

exchange information so that we have the best understanding of the legs and the ears and the, you know, tusks and the tail and what have you. So I mean, I think that's what these standards are all about.

The GSC provides an important strategic opportunity for aligning the interests of the various stakeholders, in order to ensure that all of the pieces fit together to make a "whole elephant." The GSC in turn is working not only on technical standards for data representation, but is also working to build social and organizational relationships to support creation and use of standardized genomic data. For example, the GSC has created a new peer-reviewed journal dedicated to genomic standards; scientists will be able to add publications to their CVs for providing standards-compliant genomic metadata (Garrity et al. 2009). This new journal functions as a way to align the GSC's interest in metadata compliance with academic systems for assigning credit and reputation.

By joining the GSC, CAMERA extends its network of relationships and increases its embeddedness. Through the GSC organization and standards, CAMERA gains indirect alignments with a significant number of other systems and organizations (see Figure 4).

6.2.3. *Landscape of databases*

CAMERA is not the only project developing sequence databases and tools for metagenomics research. While in one sense these systems are competing against

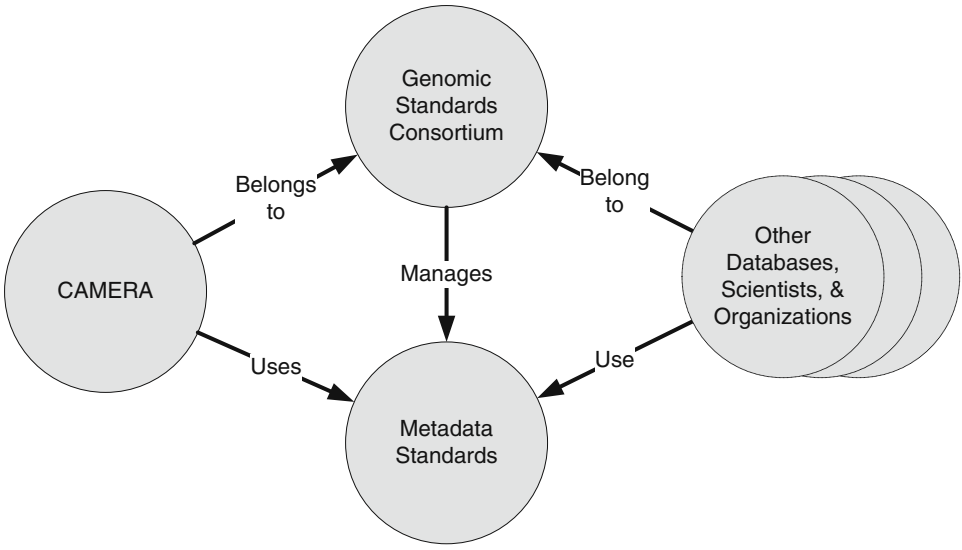

Figure 4. As our informants explained, by joining the GSC, CAMERA is able to participate in the creation and management of metadata standards, and gains indirect relationships and alignments with a significant number of other databases, scientists, and organizations.

each other (for funding, prestige, etc.), the developers see it as important that the different sequence databases operate on the same set of sequence data. One CAMERA project member told us:

> *It doesn't serve the community well if CAMERA stands out there distinguished, beating its chest, basically we have more data than [GenBank] or we have different data than [GenBank]. I would argue philosophically that's a losing strategy and CAMERA should not distinguish itself on what data it contains. Rather it distinguishes itself on the tools it provides the community to analyze and make sense of those data.*

There is a feeling in this community that a greater scientific purpose is served by having multiple systems providing different ways to look at the same data. But in order to make this possible, the CAMERA project had to create both technological and organizational alignments with other systems like GenBank. For example, it is hoped that GenBank will also accept data from the CAMERA database. The same project member went on to say:

> *So, ideally a user would come to CAMERA, upload their data, do whatever they need to do, and at some point there's a button that says submit data to GenBank.*

In order to make this ideal a reality, data exported from CAMERA must be in a format than can be understood by GenBank's computers. Furthermore, GenBank and CAMERA must agree on which data and metadata fields are required and ensure they are collected when data is submitted. There needs to be some way to make sure that CAMERA's standards for data quality are acceptable to GenBank, and that CAMERA is not submitting data that duplicates GenBank's collections. When users submit data, they are also making a legal agreement about how their data can be presented and used, so GenBank's and CAMERA's legal terms must be aligned. The simple user interface for uploading data hides a complex set of negotiations and alignments of technologies, policies, and procedures among the two organizations.

The "button" to submit data to GenBank also serves as an excellent reminder that CAMERA is an example of infrastructuring. This vision—a scientist sitting at her computer and seamlessly and transparently connecting her data into a web of global resources—drives the development of CAMERA. This linking of the local and global is the key to infrastructure:

> *An infrastructure occurs when the tension between local and global is resolved. That is, an infrastructure occurs when local practices are afforded by a larger-scale technology, which can then be used in a natural, ready-to-hand fashion. (Star and Ruhleder 1996, p. 114)*

CAMERA's developers are writing software and building fast computers. But beyond that, CAMERA represents a coordinated attempt to build socio-technical

structures that could make it possible to bring together the individual activities of thousands of scientists to produce a new kind of science. It is too early to know if CAMERA will be considered a success. Similarly, we do not know if the promise of metagenomics will be fulfilled, or if the results of this kind of "big science" will justify the costs involved with building cyberinfrastructures. But our concern in this paper is with understanding what it means to do development work with the goal of producing infrastructure, and synergizing is clearly part of that process.

In this section we have seen how, in an infrastructural context, something as seemingly straightforward as putting data into a database requires creating and managing a complex set of socio-technical relationships. Some of these activities happen mostly at the level of technology, such as building scripts that translate scientists' submitted data into to a supported format. But in order to create these technological connections, it is also necessary to leverage and align relationships within and across multiple organizational structures. The developers draw on existing relationships to create new relationships. Synergizing both depends upon and produces embeddedness: the database is always already situated in a complex, multi-dimensional web of social and technical relationships, and the work of the developers simultaneously draws on and extends those relationships and connections.

7. Discussion

One of the themes that is highlighted in the examples above is that synergizing is a complex *socio-technical* process: it is impossible to understand the development of cyberinfrastructure fully without considering both technological and social relationships. We found time and again that seemingly technical decisions were being driven by organizational or interpersonal pressures, and vice versa. A good working relationship with the developer may be more important than functionality when choosing a software component. On the other hand, an otherwise viable collaboration may be avoided because it would be too difficult to overcome technology incompatibilities. We find that explanations of how cyberinfrastructure gets built are enriched by considering how relationships between technological, organizational, interpersonal, and community concerns are managed.

We found that resource scarcity often drives synergizing. Developers are trying to be as efficient as possible and do the most with the least. Most cyberinfrastructure projects, CAMERA included, are funded by government or non-profit agencies. There is often not enough time, money, or expertise in the project to build everything internally. One software engineer on the CAMERA project told us:

> *A lot of it was quite pointedly and deliberately leveraging existing capabilities on campus as distinct from just starting up from scratch and, perhaps even worse, risking reinventing the wheel to do the same thing.*

Given this, it is often as appropriate to characterize cyberinfrastructure development as aggregation or assembly rather than creating from a blank slate.

Synergizing is also, in some respects, a strategic response to the structure of academic and non-profit research funding. Funding from agencies like the GBMF and the National Science Foundation tends to come in relatively short-term grants. These agencies also place a high priority on innovation, and structure their funding to support "transformative" and "new" work rather than ongoing support and maintenance activities. One grant administrator told us:

> *First and foremost, though, we're in the business of doing transformative grant making. And what that means is get it going, accelerate it, and then if it's really got legs, it's gonna stand on its own two feet…. So whatever we start has an end. And they're built that way on purpose.*

Some of the components that CAMERA uses began as research projects in their own right, but when their initial grant expired, they found it difficult to get new funding to support ongoing development and maintenance work. However, one of the ways that CAMERA "leverages" a technology is by hiring the developers to be part of the CAMERA team. It is important for CAMERA to have a robust system, so it essentially subsidizes continued development on these components.

Our informants frequently describe their own synergizing work not only in the context of building new relationships, but also as a way to strengthen existing relationships. Leveraging creates networks in which entities are frequently linked through multiple indirect relationships. Two organizations may be linked in a collaboration, but that collaboration is stronger when they are also using the same technologies, their staffs have good interpersonal relationships, they participate in the same consortia, they receive funding from the same sources, etc. The arrangement of relationships becomes more stable, providing more numerous and predictable opportunities for future leveraging. For example, one CAMERA developer told us:

> *I've also been involved with other projects at UCSD. So, in general what happens whenever there's a new project is they tend to look around to see if there is someone else around here who has done similar work and if they can be leveraged.*

The university environment provides a densely connected network of people who have existing professional and interpersonal relationships, have expertise in cyberinfrastructure, and operate within the same organizational hierarchy. Of course, leveraging can simplify the process of setting up relationships with other people, organizations, or technologies. Being on the same campus makes it easier to find and access potential collaborators. But at the same time, leveraging is a way to exploit existing alignments. When this developer draws on existing capabilities on campus, he knows, for example, that it will be easier to pay for services through an internal funds transfer, or to purchase identical hardware

using existing vendor contracts. Potential collaborators will be on the same academic calendar and have holidays on the same days. If the developer were to go outside the university, all of these aspects (and many more) could require new alignment work. Leveraging can reduce the scope and amount of alignment work that must be done. This is another example of embeddedness being used as a resource for development work.

Additionally, alignment in these dense and overlapping structures tends to be self-reinforcing. Sharing a metadata standard provides a reason for cyberinfrastructure projects to collaborate in the creation and maintenance of the standard. At the same time, because they are collaborating in the ongoing development of the standard, there is an obvious impetus for them to continue to use it. Because entities are already aligned via multiple organizing structures, ongoing synergizing work is easier and more productive.

7.1. Synergizing and translation

We also find that synergizing is often necessary precisely because CAMERA, its component technologies, and the science it supports are at the cutting edge of research and development (Bietz and Lee 2009). When the metagenomicists needed a database to store and share their data, they found that there was no existing database system that supported their needs. At the same time, when database developers try to determine the scientists' requirements, they discover that the community of scientists has not come to a decision about what kind of data should be in the database or what research tools are required. Metagenomics is "science in the making" (Latour 1987); the technologies and practices are being simultaneously co-developed. CAMERA's embeddedness in the developing science of metagenomics highlights that the conceptual lens of synergizing has certain similarities to the notion of translation for studying the development of cyberinfrastructure (Callon 1986; Latour 1987). One might see in the above examples instances of problematization, in GBMF requiring winners of its grants to upload their data to CAMERA; of interessement, in CAMERA's involvement with the GSC; of enrolment, in the negotiations between JCVI and CAMERA about what is done by whom; and of mobilization, in CAMERA's use of "evangelists" to convince the metagenomics community of the project's value. We find, however, that translation does not fully account for what we saw in the development of CAMERA. We describe here two specific ways in which the analytic lens of synergizing draws attention to aspects of activity not highlighted in a translational account: the process of interessement, and the notion of a center of calculation.

In a translational analysis, CAMERA is constituted via its problematization, i.e., its becoming indispensable to a variety of actors (marine microbiologists, database architects, funding agencies, gene sequence data, etc.). However, those actors are also implicated in other ways in relation to other entities. For example, many of the entities involved in CAMERA are also involved in efforts such as

GenBank. "To interest other actors is to build devices which can be placed between them and all other entities who want to define their identities" (Callon 1986, p. 208). A classic translational analysis would thus look for the ways in which CAMERA attempts to interpose itself between marine microbiologists and GenBank, or between gene sequence data and GenBank. To some extent, these sort of interpositions are present; the idea of a button in the CAMERA data interface enabling a scientist contributing gene sequence data to "Submit to GenBank" places CAMERA as an intermediary between this contributor and GenBank, as well as between the gene sequence data and GenBank. However, such a tactic does not completely dissociate either the contributor or her data from GenBank, as is implied in the notion of interessement. We argue that this example is better seen instead as an instance of synergizing, specifically as an instance of alignment. Rather than cutting off important actors (e.g., data contributors) from other entities (e.g., GenBank), CAMERA instead makes these multiple relationships productive via alignment—of data and metadata standards, of personal relationships between personnel, of organizational goals, of funding sources, etc. The analytic lens of aligning better accounts for these processes than a fully translational analysis.

CAMERA can also be seen as a center of calculation, a place where "specimens, maps, diagrams, logs, questionnaires and paper forms of all sorts are accumulated and used by scientists and engineers" in the production of knowledge (Latour 1987, p. 232). Various entities in the world, e.g., marine microbes, gene sequences, microbe sampling expedition routes, and water chemistry affecting pH, are brought to these centers by translating them to immutable combinable mobiles, e.g., tables, graphs, FASTA files, and database entries. Although framable as a center of calculation, CAMERA is not centralized. That is, the calculating does not occur within a organizational location, but rather calculation is distributed across many entities, including Calit2, the San Diego Super Computing Center (SDSC), the Scripps Institute of Oceanography (SIO), JCVI, and many others. How, then, is such a distributed center of calculation made productive? Synergizing helps account for this productivity by shifting analytic focus to the leveraging that occurs between these different components of the center. For example, personal connections between individuals at Calit2 and SDSC are leveraged so that cluster computing technologies developed at SDSC can be used in CAMERA. A translational analysis has no explicit way of accounting for such interactions, which becomes problematic, as many cyberinfrastructure endeavors fit this mold of a distributed center of calculation. Synergizing, on the other hand, draws attention to the processes by which these relationships within the distributed center are enacted and made productive.

7.2. Synergizing and articulation work

We also find that synergizing has many important similarities and differences to articulation work that are worth exploring further. Strauss distinguishes between

articulation work and articulation process. Articulation work "refers to the specifics of putting together tasks, task sequences, task clusters—even aligning larger units such as lines of work and subprojects—in the service of work flow" (Strauss 1988, p. 164). The articulation process is putting and keeping together elements of work—or as it has been most frequently described, it is the work of making sure the work gets done. In discussing articulation work, Gerson (2008) makes a distinction between the concepts of "metawork" and "local articulation."

> *Strauss used the notion of articulation work in two different senses* (e.g. *Strauss* 1988). *On the one hand, articulation work is about making sure all the various resources needed to accomplish something are in place and functioning where and when they're needed in the local situation. This means bringing together everything needed to accomplish a task at a particular time and place, including all the administrative and support functions such as janitorial services, food service, equipment maintenance, and covering for staff out sick or on vacation. The concern and emphasis in this sense are on particular situations rather than classes of activity. (Gerson* 2008*, p. 196)*

Above is what Gerson calls local articulation, the bringing together of local resources for a particular situation. The passage below illustrates what Gerson calls "metawork."

> *In its second sense, articulation work means "putting together tasks, tasks sequences, task clusters—even aligning larger units such as lines of work and subprojects in the service of work flow." In this second sense, the focus is not so much on the specifics of work in a particular local situation, as it is on making sure that different kinds of activity function together well. The two senses overlap heavily—especially when the tasks are part of the same organization and are carried out in the same place. (Gerson* 2008*, p. 196)*

If the two senses overlap heavily especially when tasks are part of the same organization and carried out at the same place, it stands to reason that the two senses overlap less when tasks are part of different organizations and carried out at different places. Indeed Strauss (1988) notes that, "Other models probably are needed to analyze the articulation process for lines of work and for encompassing organizations, as well as for interorganizational relationships" (p. 164).

Interorganizational relationships do indeed seem to be different from intraorganizational ones. Previous work on local articulation focuses on aligning tasks related to a particular location rather than classes of activity (Gerson 2008). For local articulation, the projects are located within organizations and there is a fairly well bounded pool of extant resources to draw from and bring together. Gerson describes coordination mechanisms as being concerned primarily with metawork, tasks dedicated to coordinating other tasks. At the same time both metawork and local articulation focus on modifying a "common field of work" which is the collective of things upon which an ensemble is enacting state

changes (Schmidt & Simone 1996). Synergizing differs from local articulation and metawork in that synergizing is concerned not with modifying and coordinating an existing common field of work, but with *creating the field of work itself*. Synergizing is the business of building and maintaining a common field of work and working towards the creation of a particular situation in which local articulation and metawork, including coordination mechanisms, can be enacted. This is not to say that synergizing is altogether different from metawork or local articulation. The synergizing concept of aligning is similar to the notion of metawork in that tasks and lines of work must be brought together.

As Schmidt and Simone (1996) noted previously, the distinction between cooperative work and what Gerson calls metawork is recursive: an established metawork articulation arrangement may itself be subjected to a cooperative effort of re-arrangement that may in turn need to be articulated through metawork, and so on. Just as Schmidt and Simone describe cooperative work and metawork as being recursive, we also recognize that the infrastructural work of synergizing may become recursive with metawork as the infrastructure develops. Creating the alignments necessary to share genetic data across analysis platforms may require metawork to decide how the work of alignment is divided and accomplished or what standards will be used. This metawork can become the target of further synergizing as new infrastructures are created, for example, to support standards development, and so on. Recursion in local articulation, metawork, and synergizing is worthy of further research.

Synergizing also draws attention to the difficulty of defining the "common field of work" for infrastructure development. Strauss (1988) discusses the difficult analytical problem of dealing with projects within an organization that have subprojects. In our study of the development of CAMERA, we have found another facet of complexity in that projects and sub-projects do not fit neatly into a hierarchy. Infrastructures can have components that function simultaneously as sub-projects and independent entities. Components may begin to function long before the overall infrastructure is complete, or the same component may be part of multiple infrastructures. Boundaries among projects and sub-projects are likely to be amorphous and in constant flux. Synergizing is a strategy for managing this complexity to create and maintain the common field of work.

We are also looking at an organization that includes as part of its mission to create interorganizational projects for which constituent organizational structures do not necessarily "nest" within each other or fit a layer cake model. As noted in earlier work on human infrastructure (Lee et al. 2006), the human infrastructure of cyberinfrastructure holds many forms at once (groups, organizations, networks, etc.) that are all salient and important for making infrastructure happen. As Strauss (1988) mentioned in relation to articulation processes, everyone at every level can contribute, so too with synergizing do people at various places across the various organizational hierarchies, from foundation directors to programmers, contribute to synergizing.

During the process of creating infrastructure, whole groups, software tools, communities, organizations, and other elements of human infrastructure may come and go as the project matures. A challenge for synergizing is to restrain complexity within an existing, common field of work (as per metawork and coordination mechanisms), and to also bring together elements in order to create and maintain the common field of work itself. Synergizing is a strategy for creating, managing, and utilizing complex interdependences in an embedded infrastructure that brings together multiple organizations, projects, people, and technologies.

8. Conclusion

To reiterate, it is not so much synergy that interests us, but rather the accomplishment of the process of synergizing. An infrastructure for pushing the boundaries of science is burdened with trying to match and anticipate scientific requirements. Therefore while synergizing occurs constantly, synergy itself is a moving target and an ideal state. The concept of synergizing is a way to frame the work of collaborative infrastructure development in which human infrastructure and the technical elements of infrastructure are brought into alignment in ways that will result in a larger combined effect—the ultimate goal of infrastructure. The achievement of an electricity infrastructure is far more than an aggregate of electrical generators, power lines, and wall outlets; rather it is a shift in the realm of possibility. Similarly, the groups, servers, organizations, and tools of CAMERA will result in a shift of possibility for the science of metagenomics and beyond.

In the case of building cyberinfrastructure, collaboration often occurs across a spectrum of human infrastructure, including organizations, groups, networks, teams, and other collaborative structures. This human infrastructure is brought to bear against a spectrum of technical components ranging from scripts, to web applications, to middleware, to computer clusters. Just as the human infrastructure of cyberinfrastructure (Lee et al. 2006) holds many forms at once and shifts dynamically in the process of work, the technical infrastructure also shifts dynamically as work is accomplished. Understanding cyberinfrastructure requires an approach that considers arrangements of both social and technical relationships together. Cyberinfrastructures are built in the interaction of the social and technical, and synergizing often moves across multiple social and technical structures.

This multiplicity means that a one-dimensional analysis cannot capture the work of creating cyberinfrastructure. A purely organizational analysis of configurations between GBMF, Calit2, and JCVI would likely downplay the interpersonal relationships that support those organizational connections. Similarly, focusing only on the technical work of developing a database schema would not effectively highlight the scientific process of developing genomic metadata

standards. Attending too intently to any single category of relationships—interpersonal, technical, organizational, etc.—unnecessarily constrains the analysis and limits the researcher's ability to understand how these various complex entities interact. The concept of synergizing not only captures the fact that infrastructuring consists of highly multidimensional processes, but it draws attention to the complexities of the work required to create and maintain productive relationships.

This paper extends our understandings of cyberinfrastructure development in key areas. In many ways, this paper is an expansion of the concept of the Human Infrastructure of Cyberinfrastructure (Lee et al. 2006). As human infrastructure draws attention to the diversity of collaborative structures that come into play in an infrastructural endeavor, synergizing draws attention to the socio-technical collaborations that are important for the success of infrastructure building. Synergizing is a key mechanism by which the social and technological aspects of infrastructures are connected. Synergizing highlights the work that goes into building and maintaining cyberinfrastructure.

Others have pointed to the tension between "emergence" and "intention" as being a key challenge in infrastructure development (Edwards et al. 2007; Ribes and Finholt 2007). The synergizing lens allows us to see how they need not be mutually exclusive. Because so much of the work of cyberinfrastructure development involves leveraging and aligning networks of relationships, developers are involved in ongoing decisions about with *whom* (or with *which entities*) to interact (leveraging), and *how* those relationships will work (aligning). GBMF, for example, is more concerned with whom the infrastructure will serve and how CAMERA will relate to other infrastructures than it is with exactly what the CAMERA artifacts will look like. At the same time, developers of cyberinfrastructure also have to manage intentionality from multiple sources and directions. This multiplicity of stakeholders is a key feature of infrastructures. The properties of an infrastructure emerge from the aggregation of multiple, ongoing synergizing-related decisions. Emergence is not accidental, but perhaps unpredictable because of the variety and complexity of intentions.

Finally, leveraging and aligning highlight the dual nature of infrastructural embeddedness. On the one hand, the network of relationships constrains developers' ability to design the system to match their intentions—make too big a change, and the relationships will fall out of alignment and no longer be productive. On the other hand, the network of relationships is a resource for developers—they leverage existing relationships to develop new and stronger relationships. Synergizing lets us see how these two aspects of embeddedness play out as part of infrastructuring.

Synergizing is creating interactions or cooperation to produce a greater combined effect. Synergizing enlists a vast and shifting assortment of technologies and human infrastructure that vary according to scope and size. Interactions and relationships among entities are embedded in a socio-technical

web of relationships. In the case of CAMERA we have seen that synergizing can result in durable relationships for this cyberinfrastructure itself, but will also result in durable relationships for any number of other cyberinfrastructures. The day-to-day work of synergizing may seem small, but the ultimate net effect of this infrastructure building is to look forward and change the face of science: in this case, developing the discipline of metagenomics so that it becomes established science. Synergizing is a multidirectional, multidimensional form of collaboration to bring about changes of the importance and complexity of a cyberinfrastructure to establish and support new scientific practices and disciplines.

Acknowledgments

This paper is dedicated to the late Prof. Roberta Lamb whose inspiration and encouragement were invaluable. We thank the anonymous reviewers for their work to help us improve this paper. Special thanks to the study participants who were so generous with their time. This research was supported by NSF awards IIS-0712994 and OCI-0838601.

References

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messerschmitt, D. G., et al. (2003). *Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure*. Washington, D.C.: National Science Foundation.

Berman, F. (2001). The human side of cyberinfrastructure. *EnVision, 17*(2), 1.

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

Blumer, H. (1969). *Symbolic Interactionism: Perspective and Method*. Englewood Cliffs, NJ: Prentice-Hall.

Callon, M. (1986). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St. Brieuc Bay. In J. Law (Ed.), *Power, Action and Belief: A New Sociology of Knowledge?* (pp. 196-223). London: Routledge.

Corbin, J., & Strauss, A. (2008). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (3rd ed.). Thousand Oaks, CA: Sage.

Davidson, C. N., & Goldberg, D. T. (2004). A manifesto for the humanities in a technological age. *The Chronicle Review, 50*(23), B7.

Edwards, P. N., Jackson, S. J., Bowker, G. C., & Knobel, C. P. (2007). *Understanding infrastructure: Dynamics, tensions, and design*. Ann Arbor, MI: Deep Blue.

Emerson, R. M., Fretz, R. I., & Shaw, L. L. (1995). *Writing Ethnographic Fieldnotes*. Chicago, IL: University of Chicago Press.

Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., et al. (2008). The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology, 26*(5), 541–547.

Finholt, T. A. (2002). Collaboratories. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (Vol. 36, pp. 73–107). Medford, NJ: Information Today Publishers.

Gad, C., & Bruun Jensen, C. (2010). On the Consequences of Post-ANT. *Science, Technology, & Human Values, 35*(1), 55–80.

Galison, P. (1997). *Image and logic: A material culture of microphysics*. Chicago: University of Chicago Press.

Garrity, G., Field, D., & Kyrpides, N. (2009). Standards in Genomic Sciences. *Standards in Genomic Sciences, 1*(1), 1–2.

Gerson, E. M. (2008). Reach, bracket, and the limits of rationalized coordination: Some challenges for CSCW. In M. S. Ackerman, C. A. Halverson, T. Erickson, & W. A. Kellogg (Eds.), *Resources, Co-Evolution and Artifacts: Theory in CSCW* (pp. 193–220). London: Springer.

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., & Trow, M. (1994). *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*. London: Sage.

Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New York: Aldine de Gruyter.

Granovetter, M. (1985). Economic Action and Social Structure: The Problem of Embeddedness. *The American Journal of Sociology, 91*(3), 481–510.

Hey, T., & Trefethen, A. (2003). e-Science and its implications. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 361* (1809), 1809-1825.

Hey, T., & Trefethen, A. E. (2005). Cyberinfrastructure for e-Science. *Science, 308*(5723), 817–821.

Jones, G. R., & George, J. M. (1998). The experience and evolution of trust: Implications for cooperation and teamwork. *Academy of Management Review, 23*(3), 531–546.

Karasti, H., & Baker, K. S. (2004). Infrastructuring for the long-term: Ecological information management. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 1* (Vol. 1, pp. 10020c). Los Alamitos, CA: IEEE Computer Society.

Katz, M. L., & Shapiro, C. (1985). Network Externalities, Competition, and Compatibility. *The American Economic Review, 75*(3), 424–440.

Latour, B. (1987). *Science in Action*. Cambridge, MA: Harvard University Press.

Latour, B. (1993). Ethnography of a "high-tech" case: About Aramis. In P. Lemonnier (Ed.), *Technological Choices: Transformation in Material Culture Since the Neolithic* (pp. 372–398). London: Routledge.

Lee, C. P., Dourish, P., & Mark, G. (2006). The human infrastructure of cyberinfrastructure. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work* (pp. 483–492). New York: ACM.

Marshall, E. (2001). Bermuda Rules: Community Spirit, With Teeth. *Science, 291*(5507), 1192.

Nardi, B. A., Whittaker, S., & Schwarz, H. (2002). NetWORKers and their activity in intensional networks. *Computer Supported Cooperative Work, 11*(1–2), 205–242.

National Center for Biotechnology Information (2008). GenBank Overview. http://www.ncbi.nlm.nih.gov/Genbank/index.html. Accessed 23 February 2009.

National Center for Biotechnology Information (n.d.). FASTA Format Description. http://www.ncbi.nlm.nih.gov/blast/fasta.shtml. Accessed 14 April 2009.

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

Olson, G. M., Zimmerman, A., & Bos, N. (Eds.). (2008). *Scientific Collaboration on the Internet*. Cambridge, MA: MIT Press.

The Oxford English Dictionary. (1989) (2nd ed.). Oxford, UK: Oxford University Press.

Palmer, C. L. (2001). *Work at the Boundaries of Science: Information and the Interdisciplinary Research Process*. Boston: Kluwer.

Pinch, T. J., & Bijker, W. E. (1984). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science, 14*(3), 399–441.

Pipek, V., & Wulf, V. (2009). Infrastructuring: Toward an integrated perspective on the design and use of information technology. *Journal of the Association for Information Systems, 10*(5), 447–473.

Poon, S. S., Thomas, R. C., Aragon, C. R., & Lee, B. (2008). Context-linked virtual assistants for distributed teams: An astrophysics case study. In *Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work* (pp. 361-370). New York, NY: ACM.

Ribes, D., & Finholt, T. A. (2007). Tensions across the scales: Planning infrastructure for the long-term. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work* (pp. 229-238). New York: ACM.

Schmidt, K., & Simone, C. (1996). Coordination mechanisms: Towards a conceptual foundation of CSCW systems design. *Computer Supported Cooperative Work (CSCW), 5*(2), 155–200.

Seshadri, R., Kravitz, S. A., Smarr, L., Gilna, P., & Frazier, M. (2007). CAMERA: A community resource for metagenomics. *PLoS Biology, 5*(3), e75.

Sides, S. (2007). What is CAMERA? http://camera.calit2.net/about-camera/what-is-camera. Accessed 7 March 2010.

Smarr, L. L., Chien, A. A., DeFanti, T., Leigh, J., & Papadopoulos, P. M. (2003). The OptIPuter. *Communications of the ACM, 46*(11), 58–67.

Sonnenwald, D. H. (2007). Scientific collaboration: A synthesis of collaborations and strategies. In B. Cronin (Ed.), *Annual Review of Information Science & Technology* (Vol. 41, pp. 643–681). Medford, NJ: Information Today.

Spencer, B. F., Jr., Butler, R., Ricker, K., Marcusiu, D., Finholt, T. A., Foster, I., et al. (2008). NEESgrid: Lessions learned for future cyberinfrastructure development. In G. M. Olson, A. Zimmerman, & N. Bos (Eds.), *Scientific Collaboration on the Internet* (pp. 331–347). Cambridge, MA: MIT Press.

Star, S. L. (1999). The ethnography of infrastructure. *American Behavioral Scientist, 43*(3), 377–391.

Star, S. L., & Bowker, G. C. (2002). How to infrastructure. In L. A. Lievrouw & S. Livingstone (Eds.), *The Handbook of New Media* (pp. 151–162). London: SAGE Publications.

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research, 7*(1), 111–134.

Strauss, A. (1988). The Articulation of Project Work: An Organizational Process. *The Sociological Quarterly, 29*(2), 163–178.

Traweek, S. (1988). *Beamtimes and Lifetimes: The World of High Energy Physicists*. Cambridge, MA: Harvard University Press.

Wulf, W. A. (1993). The collaboratory opportunity. *Science, 261*(5123), 854–855.