Adapting Cyberinfrastructure to New Science: Tensions and Strategies

Matthew J. Bietz
University of California, Irvine
Department of Informatics
Irvine, CA 92697-3440
+1 (949) 824-2901
mbietz@uci.edu

Charlotte P. Lee
University of Washington
Human Centered Design & Engineering
Seattle, WA 98195
+1 (206) 543-2567
cplee@uw.edu

ABSTRACT

Scientific information infrastructures, or cyberinfrastructures, are expected to operate over long time scales, but this creates challenges for the design of those infrastructures. This paper reports on a qualitative study of cyberinfrastructure development in the emerging field of metagenomics to illustrate some of the issues that can arise when cyberinfrastructures are faced with new scientific communities, practices, and research questions. New science inevitably brings new forms of data, new analysis tools, and the need to recontextualize existing data. Cyberinfrastructures must be prepared to adapt to the new scientific context. In this study, developers employed three strategies for addressing new scientific requirements: *work-arounds*, *extensions*, and *from-scratch* development. These strategies are informed by the tension between fitting the CI to the needs of a specific community and maintaining interoperability across systems.

Categories and Subject Descriptors

H.5.3. **[Information Interfaces and Presentation]**: Group and Organization Interfaces.

General Terms

Design, Management, Human Factors, Theory.

Keywords

Cyberinfrastructure, e-Science, metagenomics, sustainability, tailoring, interoperability.

1. INTRODUCTION

Cyberinfrastructures (CI) are large-scale information infrastructures based on advanced computational, networking and organizational capabilities to support distributed knowledge sharing and production. Cyberinfrastructures operate over relatively long time scales, but this creates certain challenges for the designers of these infrastructures. Often the CI is expected to persist through funding cycles, changes in technologies, the coming and going of people involved in the project, and larger social and policy changes [11,31]. One particularly difficult challenge is that as the infrastructure evolves, the user base may change. As users change their focus or new users arrive they present a new set of requirements and infrastructure needs. Here we use the emergent field of metagenomics research to illustrate some of the challenges that arise when scientists begin to use existing information infrastructures to answer new research questions.

Metagenomics, sometimes called population genomics or environmental genomics, is a "new science" that allows scientists to study the genetic composition of populations of microorganisms to understand biological diversity, microbes' functional roles, and microbial impacts on and adaptations to specific environments. Metagenomics is an interdisciplinary approach, using the analysis of genetic sequence data to answer questions in fields as diverse as environmental remediation, cancer research, drug discovery, marine microbiology, and power generation [28].

Metagenomics is enabled by new laboratory methods, advances in sequencing technologies, and cutting edge information infrastructures. In the past, geneticists and genomicists who wanted to study an organism's DNA or RNA had to isolate individual organisms and grow them in the laboratory in order to extract enough genetic material for analysis. This material would be analyzed to produce a "sequence" of letters representing the chain of amino acids. This process was slow and expensive. And it has been estimated that less than 0.1% of the world's microorganisms are amenable to culturing in the laboratory, severely limiting the species of organisms that microbiologists could study.

New techniques and technologies have been developed that make it possible to bypass the culturing step while significantly lowering the cost of DNA sequencing. These changes give scientists access to a wealth of genetic information from organisms that previously could not be studied. Not only can scientists study new organisms, they can also ask new kinds of questions about them. Since it is possible to sequence DNA without culturing the organisms, it is also no longer necessary that all the DNA come from the same organism. Genetic information can be collected with the *population* of microbes as the unit of analysis. This new field of study is sometimes described as "beyond genomics," or *meta*genomics. The complete set of genetic information collected from an environmental sample is a *metagenome*.

Here we investigate the growth of metagenomics to explore some of the ways that the emergence of a new community of scientists with new research questions and new information needs can challenge existing information infrastructures. We draw on a

iConference 2012, February 7-10, 2012, Toronto, ON, Canada.
ACM 978-1-4503-0782-6/12/02.

qualitative study of cyberinfrastructure development in metagenomics to detail three strategies that developers employ to adapt cyberinfrastructures to new science. We discuss how these strategies are informed by the tension between fitting the CI to the needs of a specific community and maintaining interoperability across systems.

2. BACKGROUND

2.1 Scientific Cyberinfrastructure

With advances in computation, networking and communication technologies, there has been a move toward a new mode of collaborative and cooperative science. This science is often conducted in a distributed fashion, with large shared data sets, computationally-intensive analysis, and large numbers of contributing scientists. This new mode of science has been given various monikers including e-Science, collaboratories, virtual science, or Big Data science [2,13,17]. This new kind of science relies on *cyberinfrastructure*, "a new class of infrastructure based upon distributed computer, information and communication technology. If infrastructure is required for an industrial economy, then we could say that cyberinfrastructure is required for a knowledge economy" [3]. Cyberinfrastructures are more than just hardware and software—they also include the groups, organizations, and networks that comprise the "human infrastructure of cyberinfrastructure" [19]. Funding organizations around the world (like the US National Science Foundation) are making significant investments in cyberinfrastructure development, motivated in part by new opportunities to conduct research that will answer some of society's most pressing environmental, health, and energy problems.

With this investment in infrastructure, there is an increasing interest in what happens to cyberinfrastructures over time. The sustainability of cyberinfrastructures is a pressing concern, especially around issues of providing resources beyond short-term project lifecycles [22,33] and maintaining data, software, and other components of the scientific record [20,21,36]. Another aspect of sustainability, however, is understanding how cyberinfrastructures keep pace with changes in the science they support. In this paper we examine this aspect of cyberinfrastructure sustainability.

2.2 CI for the Genetic Sciences

Cyberinfrastructure plays a major role in the genetic sciences, where scientists have long recognized the importance of sharing genetic sequence data. The field has strong norms around data sharing, backed up by a commitment by journal publishers not to publish analyses of genetic data unless the data is submitted to publicly accessible databases [24]. GenBank, for example, has been collecting and distributing genetic sequence data since 1982 [26]. GenBank is just one of several infrastructures that provide storage of genetic sequence data and facilities for analyzing and visualizing that data.

These collections of shared data play a more important role in the genetic sciences than they might in other fields. The databases are, to a certain extent, a representation of the state of genetic knowledge [6]. Many analyses begin by comparing a new genetic sequence to the sequences in the database in order to, for example, find out if the sequence is new or to predict the expression of proteins based on similarity to known genes. The GenBank database provides one of the most comprehensive collections of submitted data, but other databases are also used. The *RefSeq* database, for example, is a curated collection of GenBank records that removes duplicate and low-quality entries. Other databases are created to collect particular kinds of organisms (e.g. pathogens), to serve particular communities, or to support specialized analyses.

The data in these databases is submitted by the scientists who conduct the DNA sequencing and analysis. While the databases may have their own underlying architectures, data sharing among the scientists and databases is supported by a strong standard called FASTA, which specifies a uniform file format for representing sequences using individual letters to stand for amino acids [27]. Many of these systems also provide standard tools like the Basic Local Alignment Search Tool (BLAST) which allow scientists to compare new genetic sequences with those in the database [1].

Even the least specialized of these infrastructures (e.g. GenBank, which collects all publicly available sequence data) were designed to support particular kinds of data and analysis. The database may assume a class of research questions and a fundamental unit of analysis (e.g. the gene rather than the whole genome). Certain data were included in the database while others were not. The database better supports particular modes of searching. Data are stored in formats that make some kinds of analysis easier than others. Once these systems are operational, they also have a certain inertia or resistance to change. It is difficult and expensive to change the structure of a database that now holds years if not decades of data. This paper addresses the question of how cyberinfrastructures adapt when a "new science" like metagenomics emerges with a new set of requirements.

2.3 Metagenomics: A "New Science"

Metagenomics is a new development within the general field of genetic sciences [28]. Metagenomics provides an interesting case study for understanding cyberinfrastructure development in part because of the field's rapid growth. Indeed, the term was only coined in 1998 [16], and by mid-2005, nine major metagenomic sequencing projects had been completed [8]. Interest in these techniques is growing: for example, the Metagenomics 2008 conference attracted more than 250 participants, and the NIH is funding a major project to study the human microbiome [http://commonfund.nih.gov/hmp/]. A PubMed search for "metagenom*" returned 732 papers published in 2010 [25].

While certain aspects of metagenomics are radically different from earlier genetic and genomic approaches, there is also a level of continuity across the domains. Metagenomic analyses use the same sequence data that are used in other genetics-based fields, and tools like BLAST are still useful to compare new genetic sequences to sequences generated by other scientists. Much of the basic functionality provided by infrastructures like GenBank is equally useful for genetics, genomics, and metagenomics. For example, an important question in almost any metagenomic study is whether any of the new sequences have been previously identified. This requires that metagenomics researchers have access to as complete a set of genetic sequence data as possible, regardless of whether the original data came from genetic, genomic or metagenomic studies.

The community of metagenomics researchers also has significant overlaps with the genomics and genetics communities. The people involved in this research—the potential users of a metagenomics cyberinfrastructure—frequently conduct genetic, genomic, and metagenomics analyses in the same laboratory, and sometimes within the same project. Infrastructures are closely linked to particular communities-of-practice [35], and it is not clear that metagenomics represents an entirely new or distinct community of scientists.

At the same time, metagenomics and its associated laboratory techniques bring a new set of data storage and analyses requirements to existing cyberinfrastructure. One of the consequences of new DNA sampling and sequencing technologies is that DNA sequencing has become relatively inexpensive. While sequencing costs were around $10 per base pair in 1990 [30], today researchers pay a few cents per thousand base pairs. The amount of DNA sequence data being produced is overwhelming, to the extent that data storage and computation requirements are outpacing Moore's law [10].

In addition to simply having more data, metagenomics also assumes a different unit of analysis. Rather than focusing on the gene or even whole genome of an organism, metagenomicists work at the level of a community or population of microorganisms. Many existing sequence databases cannot easily represent this level of relationships among data.

This points to a fundamental tension in the development of infrastructures that Borgman called "tailoring vs. interoperability" [7]. On the one hand, cyberinfrastructures become more valuable when the design of the tools and databases have a good fit to the specific scientific questions being asked [6]. Metagenomics has unique requirements that are not met by existing genetic and genomic infrastructures, and the most useful systems will be those that are tailored to these requirements. On the other hand, metagenomics, both in content and practice, is not distinct from earlier genetic and genomic approaches, and interoperability across the fields is valuable. Finding the appropriate balance between these aspects is an ongoing challenge for the design of cyberinfrastructure. This paper examines how developers of cyberinfrastructures have addressed this challenge.

3. METHODS

This research reports on a three-year ethnographic study of the development of cyberinfrastructure to support metagenomics research. This study includes both an in-depth examination of one particular cyberinfrastructure development project, and a broad survey of information infrastructures serving metagenomics researchers. We have conducted forty-three interviews with metagenomics researchers, computer scientists, bioinformaticists, and others involved in the development of metagenomics cyberinfrastructures. We have also conducted over 100 hours of formal and informal observation, including attending development meetings, laboratory meetings, meetings of the Genomic Standards Consortium, workshops and conferences. Data were analyzed using a grounded theory approach [15]. As interview transcripts and field notes were generated they were open coded in Atlast.ti qualitative data analysis software [4,9]. Descriptive and analytic memos were written based on the coded data. Iterative coding and memoing continued as new data came in.

4. FINDINGS: NEW QUESTIONS FOR OLD INFRASTRUCTURES

This study provides insight into the work of cyberinfrastructure development in the face of changing science. While there were numerous aspects of cyberinfrastructure that were affected (ranging from new funding sources to new ethical and privacy concerns), we focus on key aspects of how cyberinfrastructures dealt with data. This section details two problem areas with which CI developers in our study struggled. The first is that metagenomics uses new kinds of data (and associated tools) that must be supported in the cyberinfrastructure. The second has to do with how metagenomics sheds new light on legacy data. The final part of this section outlines different strategies that developers of metagenomics cyberinfrastructures have employed to adapt to the new science.

4.1 New Data

As discussed above, metagenomics researchers work with new kinds of data at new levels of analysis. Developers of metagenomics cyberinfrastructures struggled with how to provide the support needed for this new data. A CI development project leader told us:

> These technologies are further driven by metagenomics; radically driven. Those need new software tools to help manage the data. That's the database challenge.

It is a "challenge" for the designers of databases to create systems that can usefully store, organize, retrieve, and analyze metagenomics data.

One of the reasons this is a problem for metagenomics developers is because the GenBank "Flat File Format" had become the *de facto* standard for storing and transferring sequence data. When genetic data are submitted to GenBank, the genetic sequence is represented by a series of letters using the FASTA format. However, the full record includes various other information, including the accession number and other administrative fields, citations of the original publication associated with this data, and a "feature table" that describes certain characteristics of the genetic sequence that was submitted. (A sample GenBank record can be seen at http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html). "Features" include information like the species name of the organism, the chromosome number for the sequence, and various information about identified genes and their function. This information was sufficient for the kinds of genetic research being conducted when GenBank was new.

However, one of the key focus areas in metagenomics is the relationship between microbes and their environments. For example, there is great interest in simply characterizing how microbial populations differ around the world [32]. But there are also a set of questions about the microbes that exist in highly specialized environments, like acidic mine runoff [5] or the termite gut [38]. While the metagenomics data is still represented by the same FASTA format that a geneticist or genomicist might use, there are important differences as well. First, whereas the GenBank record assumes the sequence as the unit of analysis, metagenomic data typically involves many sequences representing the population of organisms. In other words, the metagenomicist wants to study not only the sequence itself, but also how the sequence is related to a number of other sequences from the same sample.

Second, studying how the environment affects microbial life (and vice versa) requires scientists to collect contextual "metadata" that describe where the samples were found. One metagenomics researcher explained why metadata were important for his research:

> When I refer to metadata, I generally refer to ecological variables associated with the time that the samples that we use to produce that sequencing data were collected…. Metadata is vitally important for us in order to ascertain the ecology. Otherwise, it's just a bunch of sequences and

> you're shooting in the dark because you've got nothing to tie it to the real world.

Metadata allow the scientist to understand the environment from which the sample was collected. The metagenomicist might collect GPS coordinates, temperature, pH, time and date, etc., and attach these metadata to the sequence data. Most genetic and genomic databases use some version of the GenBank format, which was not designed to handle data this complex. This format makes it difficult to link sequences together in a way that represents a "population" of organisms, and it does not make provisions for environmental metadata.

Along with this new data, scientists need new tools to search, analyze and visualize the data. For example, a common question in marine metagenomics involves understanding how ocean temperature affects the diversity of the local microbiome. Not only would this require temperature data, but also the ability to query it, include it in analyses, and create visualizations around it. This kind of question would be almost impossible to answer with the data structures and tools provided in cyberinfrastructures created for traditional genetics and genomics researchers.

4.2 Recontextualizing Existing Data

Metagenomicists bring new data to existing infrastructures, but they also want to ask their new questions about old data. Often to ask a new question requires putting the old data into the new metagenomic context. For example, even if contextual metadata were not stored in the database originally, there may be sources (like the publication record) that could be used to populate new fields in the database. However, reformatting data or retrospectively adding metadata are expensive tasks, especially when the work may need to be done again for the next group of scientists who pose a new question.

Another issue arises in that new metagenomic data may change the interpretation of legacy data. For example, many analyses of genetic data (e.g. understanding evolutionary similarity, or predicting gene function) involve comparing new sequence data to the existing database. However, as new data are entered into the database, it is important to re-analyze the existing data against the new, larger set to improve the quality of the analyses. As metagenomic data is added at a phenomenal rate to these databases, the computational problems are becoming immense. One database developer told us about the difficulty of reanalyzing existing data:

> So you do need to go back from time to time and do all [the analyses] from scratch…. So the problem there is that we need to do periodic updates and periodic updates are every three months…. Now if new data is coming at an increasing pace, we are already at the point where even really big infrastructures and big computer clusters cannot really support all that.

As the size of the database grows, it takes increasing computational capacity to simply reanalyze old data, let alone work on the new data.

Beyond these issues of computational power, scientists are also refining and expanding theory. In genetics and genomics, for example, scientists are finding that some prior assumptions about how genes operate were mistaken (e.g., the role of "non-coding" regions of DNA), necessitating a reconsideration of old data and interpretations. This reconsideration often requires modifying existing tools and databases to fit the new theory.

4.3 Infrastructure Adaptation

Among the metagenomics cyberinfrastructures that were considered in this study, three different strategies emerged through which infrastructures were adapted to the new science. We call these strategies *work-arounds*, *extensions*, and *from scratch*. This section illustrates each of these approaches by examining how four different cyberinfrastructures serving the genetic sciences have dealt with new metagenomic data.

One strategy that has been adopted has been to create *work-arounds* for existing infrastructures to adapt them to new uses and questions. A work-around is an adaptation of practices or technologies within an infrastructure that meets the needs of a particular local context or subset of users, but is not fully sanctioned or supported at the infrastructural level. GenBank provides an example of a work-around to deal with contextual metadata for metagenomics data. As discussed above, GenBank does not provide much support for metadata about the environment from which a metagenome was collected. The Genomic Standards Consortium (GSC) has published a checklist of contextual metadata fields that should be specified for any metagenomics dataset [12]. However, the GenBank feature table does not have fields that match the metadata categories described by the GSC's standards. Metagenomicists want GenBank to support contextual metadata, but it is not a simple matter to make changes to the GenBank architecture. The database has decades of genetic data in the current format. Countless scientists have written analysis tools that may break if GenBank changes. But beyond the technological constraints, GenBank is also organizationally constrained. GenBank is a member of the International Nucleotide Sequence Database Collaboration (INSDC), a collaborative partnership among GenBank, the DNA Data Bank of Japan, and the European Molecular Biology Laboratory. These three institutions have agreed to a common database format and maintain synchronized databases. Changing the feature tables requires an agreement by their International Advisory Committee that publishes the data definitions. At a recent meeting of the GSC, GenBank and EMBL representatives wanted to find a way to support contextual metadata but felt that getting fields added to the feature table could take years and might not be approved by the Advisory Committee. As an alternative, the developers created a work-around.

This work-around adapts an existing free-text comment field to hold important metadata in a "structured comment" that uses text formatting to mimic a table of fields and data [14]. While the metadata are not given their own fields in the database, this work-around provides the benefit that it can be used immediately without disrupting the existing infrastructure. Metagenomics researchers who have contextual metadata can store it in GenBank records without affecting how other geneticists or genomicists use the system. On the other hand, because work-arounds are not supported within the infrastructure, they can be difficult to use, lack standardization, and do not provide full integration of the new science. Here, for example, while the metadata will be stored, they will not be indexed or as searchable as they might otherwise be, and they will not be subjected to GenBank's error checking and validation processes. However, work-arounds can be useful interim steps toward the kind of infrastructural adaptations discussed next.

A second approach involves *extending* existing infrastructures to support metagenomic data. Like work-arounds, extensions to infrastructure involve the addition of new capabilities without significant alteration to existing capabilities. However, extensions are supported at the infrastructural level, and allow new uses of the infrastructure without disrupting current users. Two cyberinfrastructures, IMG [23] and The SEED [29], have extended their systems to include new metagenomics tools and support for metagenomics data. IMG and The SEED were both originally designed to store and analyze genomic data. For example, The SEED provided an analysis facility called Rapid Annotation using Subsystem Technology (RAST) that identifies and characterizes the various subunits of a genomic sequence. In both cases, developers have *extended* these systems to support the new requirements of metagenomics data and analysis. IMG and RAST are still targeted at genomic analysis, but both systems now have separate interfaces for newer metagenomic tools (called IMG/M and MG-RAST). Even though these are new capabilities, the metagenomics tools are built on the same underlying technologies as the genomic systems. Extending infrastructures in this way requires significantly more effort than work-arounds, as it usually involves both some reconfiguration of the existing infrastructure and creating new capabilities. Even though extensions are supported within the infrastructure, there may still be difficulties in reconciling the new science with legacy systems. However, in these cases, extending the infrastructures allowed the developers to realize efficiencies by building on an installed base, and they were able to provide new capabilities without disrupting the existing system. Both IMG and The SEED are smaller infrastructures than GenBank and have less technological and organizational inertia from their legacy systems. As a result, they are able to officially incorporate and support the new science through extending their infrastructures.

A third approach, taken by projects like the Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) [34], creates new infrastructure *from scratch* specifically to support the new science. CAMERA is a cyberinfrastructure project begun in 2006 and intended to provide resources for high-volume data storage and analysis in metagenomics. CAMERA was commissioned by the Gordon and Betty Moore Foundation in response to scientists' concerns that "the existing databases are simply not capable of providing us with the ability to do what we need to do with these data" (CAMERA developer). While CAMERA incorporated some existing tools, the whole system (database, analysis platform, and human infrastructure) was a newly created infrastructure. Developing infrastructure from scratch means creating a new configuration of people and technologies to support the new science. This approach allows the developers to tailor the systems to the community's specific needs without having to worry about maintaining legacy systems or supporting other communities. For example, support for contextual metadata was a key requirement for the database structure from the beginning rather than being added in as an extension or work-around. This can also be a very expensive option in terms of both time and money, and can make it more difficult to use legacy data and tools. Having a separate infrastructure for a new science may also reinforce a separation between scientific communities that may benefit from greater interaction. However, from-scratch development may ultimately provide a better fit to the demands of the new science.

5. DISCUSSION

Cyberinfrastructures that serve metagenomics are facing challenges including new kinds of data and the need to recontextualize legacy data. Developers have addressed these challenges by creating work-arounds, extending existing cyberinfrastructures, or creating new infrastructures from scratch. This section turns to a discussion of two sets of concerns around the development and maintenance of cyberinfrastructures. First, we consider the issue of preparing for scientific change. Then we return to a discussion of finding a balance between tailoring and interoperability.

5.1 Preparing for Scientific Change

While this paper has only dealt with the specific example of metagenomics, all scientific domains can be expected to change over time [18]. As such, cyberinfrastructures should be prepared to deal with new and changing requirements from their user communities. If we accept the history told by some metagenomics researchers, metagenomics is a "logical progression" from genetics and genomics, and these future needs could have been predicted.

> The concept was simple: Take seawater and capture all the microorganisms swimming in it on filters with microscopic pores, isolate the DNA from all the captured organisms simultaneously.... Rather than focusing on the hunt for one particular type of life, we would obtain a snapshot of the microbial diversity in a single drop of seawater-a genome of the ocean itself. This was, to me, *a straightforward extension* of work that had started with the EST method and led to the whole-genome shotgun approach, then the first genome of an organism in history, and then of course to the human genome. [37, emphasis added]

Here, one of the pioneers of metagenomics writes about the new science as if it were simply the next step in the march of progress. While this version of the origin of metagenomics creates a compelling narrative, it does not recognize two important aspects of these scientific changes.

First, as science has "progressed" through these phases, it has not left old questions behind. There are still scientists who are studying the functions of individual genes, and there are still scientists who are studying the genomes of individual organisms. Metagenomics has not supplanted these fields. In fact, it is essential for metagenomicists that research continues in genetics and genomics:

> It would help us tremendously in doing metagenomics if we had a wide range of reference genomes.... The NIH is funding 400 complete genomes of microbes that live in humans. And again, these are to give us standards and to allow us to interpret metagenomic data more rigorously. So first of all, as far as I'm concerned, we've only begun to sequence. We need to sequence - whole genome studies need to go on to expand the opportunities in studying evolution and getting many specific genes and models for human disease and for understanding biology.

Metagenomics actually makes genetic and genomic studies more important, not less. Additionally, new metagenomic techniques are also changing the way geneticists and genomicists do their work. For example, shotgun sequencing not only allows for the sequencing of populations of microorganisms, it also makes it possible to sequence genomes from organisms that could not be cultured in a laboratory. Maintaining cyberinfrastructure's relevance is more than simply a matter of following a scientific trend. In this case, metagenomics adds new requirements without taking any away. This suggests a larger scope for cyberinfrastructure, and consequentially the need to find additional resources to support the additional breadth.

Second, while the quotation above suggests that metagenomics is a "straightforward extension" of earlier science, the progression from genetics through genomics to metagenomics is logical only in retrospect. The development of metagenomics was by no means a foregone conclusion, and scientists found that they had to work hard to convince their peers that these techniques were valid. One scientist explained that peer reviewers from genetics and genomics were critical of her early metagenomics articles:

> Not only has there been this distrust between the two fields, the genomics and the traditional fields—I think it's becoming more acceptable—but now metagenomics has come in too. So we're not just talking about sequencing entire genomes, we're talking about populations of genomes and defining what's there based solely on sequence similarities to those genomes. So what I've - I'm taking a huge leap here. I'm saying I have these 50,000 sequences. They're very distantly related to these sequences from [other] genomes. I know nothing about their physiology. I don't know what they infect. I don't know their reproductive lifecycle. I don't know anything about them. I'm just giving them a name based on the history of those sequences. So I think I'm taking an even farther leap.... And I think we try not to tread too heavily upon people's toes. We don't want people to think we're trying to take over their fields and that these approaches are the end all to the field.

Traditional approaches to identifying microbes rely on direct examination of microbes' physiology, pathogenesis, and reproduction. The adoption of metagenomic techniques was controversial, and this scientist found that using only metagenomic techniques was not readily accepted by peer reviewers. This new way of looking at microbial populations was not predicted by early geneticists and genomicists, the science is not without its detractors, and it is not entirely clear how these techniques will unfold into the future.

These observations highlight significant challenges for the development and maintenance of cyberinfrastructure. As science changes over time, scientists will need different things from cyberinfrastructures. While some research questions will persist, others will change and new research questions will be asked. A new science like metagenomics brings new questions and new communities of scholars with different ways of understanding the world. The requirements for information infrastructures develop and change as the science and communities change. Just as it is impossible to predict with any certainty how a scientific field will develop, it is equally impossible to predict all future information infrastructure requirements. As such, sustainability will be less a function of how well cyberinfrastructures anticipate future needs, and more a function of how well they can adapt to new science.

5.2 Tailoring vs. Interoperability

As discussed in §2.3, there is a fundamental tension between tailoring an infrastructure to a specific community and ensuring interoperability across multiple systems [7]. Genomic cyberinfrastructures have adapted to new metagenomics science through work-arounds, extensions, and from-scratch development. In the examples presented here, the strategy chosen for each cyberinfrastructure reflects the weights given to each end of this tension. In other words, maintaining the interoperability of the cyberinfrastructure requires a certain inertia or resistance to change. Maintaining a good fit between the needs of the scientist users and the capabilities of the infrastructure, however, necessitates making changes to keep up with the shifting nature of the science.

The four cyberinfrastructures discussed above have each prioritized interoperability and tailoring in different ways. GenBank clearly favors interoperability. It has a significant history that includes large amounts of collected data, a huge user base, and strong organizational commitments. It's mission prioritizes breadth of data collection over the fit to specific scientific communities. The GenBank format has become a standard for data representation in the genetic sciences. For GenBank, the risk of potentially disrupting current practices and arrangements is greater than the benefit of incorporating the new scientific needs. In this case, the work-around, which minimizes the potential for disruption, is the appropriate strategy.

CAMERA emerges from the opposite end of the tension. This from-scratch project came about because a community of scientists wanted tools that specifically fit their needs. The price of this fit was not just the monetary expense but also, at least at first, having to deal with an infrastructure that was not as robust, changed frequently, and required learning the new systems.

IMG and The SEED have chosen to more equally balance tailoring and interoperability by extending their cyberinfrastructures. Because they maintain interoperability in the capabilities and connection to their legacy systems, they may not be able to provide the close fit to the new science that comes with from-scratch development. On the other hand, by incorporating the new requirements into the infrastructure to a greater degree than work-arounds, extending the infrastructure provides relatively good fit to the new science.

None of these strategies is inherently better than the others. Instead, it is important to consider which strategy is most appropriate for the given situation. This decision must be based on consideration of a number of factors including the history of the current infrastructure; the importance of maintaining links to legacy data, tools, and systems; the cost of potentially disrupting current users of the infrastructure by making changes; and the flexibility of both the technological and human infrastructures.

6. CONCLUSION

Scientific information infrastructures that persist over long time scales must respond to the emergence of new science. New science brings with it a new set of research questions, data, tools, scientific communities, and ways of understanding legacy data. The introduction of metagenomic approaches in molecular biology highlights the dynamic nature of both the human and technological aspects of cyberinfrastructure. Developers must manage the evolution of cyberinfrastructures in response to changing user needs and requirements. Work-arounds provide a way for cyberinfrastructures to minimally adapt to new science without risking the coherence or interoperability of the infrastructure. From-scratch development can provide a more tailored system, but often at risk of disrupting existing practices and systems. Extending existing cyberinfrastructures provides a way to more evenly balance new requirements with existing uses. Over time new scientific communities and practices will develop. These strategies provide ways for cyberinfrastructures to adapt to the changing scientific context.

7. ACKNOWLEDGMENTS

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

8. REFERENCES

[1] Altschul, S. F., Gish, W., Miller, W., et al. Basic local alignment search tool. *Journal of Molecular Biology, 215*, 3 (1990), 403-410. DOI= http://dx.doi.org/10.1016/S0022-2836(05)80360-2.

[2] Aranova, E., Baker, K. S., & Oreskes, N. Big Science and Big Data in biology: From the International Geophysical Year through the International Biological Program to the Long Term Ecological Research (LTER) Network, 1957-Present. *Historical Studies in the Natural Sciences, 40*, 2 (2010), 183-224. DOI= http://dx.doi.org/10.1525/hsns.2010.40.2.183.

[3] Atkins, D. E., Droegemeier, K. K., Feldman, S. I., et al. *Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure*. National Science Foundation, Washington, D.C., 2003.

[4] Atlas.ti v. 6. (2011). [computer software]. Berlin: ATLAS.ti Scientific Software Development GmbH.

[5] Baker, B. J., Tyson, G. W., Webb, R. I., et al. Lineages of acidophilic archaea revealed by community genomic analysis. *Science, 314*, 5807 (2006), 1933-1935. DOI= http://dx.doi.org/10.1126/science.1132690.

[6] Bietz, M. J., & Lee, C. P. Collaboration in metagenomics: Sequence databases and the organization of scientific work. In *Proc. ECSCW 2009*, Springer-Verlag (2009), 243-262.

[7] Borgman, C. L. *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*. MIT Press, Cambridge, MA, 2000.

[8] Chen, K., & Pachter, L. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput Biol, 1*, 2 (2005), 106-112. DOI= http://dx.doi.org/10.1371/journal.pcbi.0010024.

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

[10] Dooling, D. (2009). Maximizing utility of genome sequence data, *The Biology of Genomes*. Cold Spring Harbor, NY.

[11] Edwards, P. N., Jackson, S. J., Bowker, G. C., et al. (2007). *Understanding infrastructure: Dynamics, tensions, and design*. Ann Arbor, MI: Deep Blue.

[12] Field, D., Garrity, G., Gray, T., et al. The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology, 26*, 5 (2008), 541-547. DOI= http://dx.doi.org/10.1038/nbt1360.

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

[14] Genomic Standards Consortium. (2009). *MIGS/MIMS Structured Comment*, [Web Page]. http://gensc.org/gc_wiki/index.php/MIGS/MIMS_Structured_Comment [2009, 2 Dec.].

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

[16] Handelsman, J., Rondon, M. R., Brady, S. F., et al. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol, 5*, 10 (1998), R245-249. DOI= http://dx.doi.org/S1074-5521(98)90108-9.

[17] Hey, T., & Trefethen, A. e-Science and its implications. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 361*, 1809 (2003), 1809-1825. DOI= http://dx.doi.org/10.1098/rsta.2003.1224.

[18] Kuhn, T. S. *The structure of scientific revolutions* (Second Edition ed.). University of Chicago Press, Chicago, 1962.

[19] Lee, C. P., Dourish, P., & Mark, G. The human infrastructure of cyberinfrastructure. In *Proc. CSCW 2006*. ACM Press, 2006, 483 - 492. DOI= http://dx.doi.org/10.1145/1180875.1180950.

[20] Lee, J. W., Zhang, J., Zimmerman, A. S., et al. DataNet: An emerging cyberinfrastructure for sharing, reusing and preserving digital data for scientific discovery and learning. *AIChE Journal, 55*, 11 (2009), 2757-2764. DOI= http://dx.doi.org/10.1002/aic.12085.

[21] Lynch, C. Big data: How do your data grow? *Nature, 455*, 7209 (2008), 28-29. DOI= http://dx.doi.org/10.1038/455028a.

[22] Mackie, C. J. Cyberinfrastructure, institutions and sustainability. *First Monday, 12*, 6 (2007).

[23] Markowitz, V. M., Ivanova, N., Szeto, E., et al. IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Research, 36* (2008), D534-D538. DOI= http://dx.doi.org/10.1093/nar/gkm869.

[24] Marshall, E. Bermuda Rules: Community Spirit, With Teeth. *Science, 291*, 5507 (2001), 1192. DOI= http://dx.doi.org/10.1126/science.291.5507.1192.

[25] National Center for Biotechnology Information. (April 30, 2011). *PubMed*, [Online database]. US National Library of Medicine. http://www.ncbi.nlm.nih.gov/pubmed/.

[26] National Center for Biotechnology Information. (2008). *GenBank Overview*, [web page]. http://www.ncbi.nlm.nih.gov/Genbank/index.html [2009, 23 February].

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

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

[29] Overbeek, R., Begley, T., Butler, R., et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research, 33*, 17 (2005), 5691-5702. DOI= http://dx.doi.org/10.1093/nar/gki866.

[30] Powledge, T. M. How many genomes are enough? *The Scientist, 4*, 1 (2003).

[31] Ribes, D., & Finholt, T. A. Tensions across the scales: Planning infrastructure for the long-term. In *Proceedings of*

the 2007 International ACM Conference on Supporting Group Work. ACM, New York, 2007, 229-238. DOI= http://dx.doi.org/10.1145/1316624.1316659.

[32] Rusch, D. B., Halpern, A. L., Sutton, G., et al. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biology, 5*, 3 (2007), e77. DOI= http://dx.doi.org/10.1371/journal.pbio.0050077.

[33] Schopf, J. M. Sustainability and the Office of CyberInfrastructure. In *Proc. Eighth IEEE International Symposium on Network Computing and Applications, 2009.*, IEEE Computer Society (2009), 1-3.

[34] Seshadri, R., Kravitz, S. A., Smarr, L., et al. CAMERA: A community resource for metagenomics. *PLoS Biology, 5*, 3 (2007), e75. DOI= http://dx.doi.org/10.1371/journal.pbio.0050075.

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

[36] Stewart, C. A., Almes, G. T., & Wheeler, B. C. *Cyberinfrastructure Software Sustainability and Reusability: Report from an NSF-funded workshop*. Indiana University, Bloomington, IN, 2010.

[37] Venter, J. C. *A Life Decoded: My Genome: My Life*. Viking, New York, 2007.

[38] Warnecke, F., Luginbuhl, P., Ivanova, N., et al. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature, 450*, 7169 (2007), 560-565. DOI= http://dx.doi.org/10.1038/nature06269.