I. Wagner, H. Tellioglu, E. Balka, C. Simone, and L. Ciolfi (eds.).
ECSCW'09: Proceedings of the 11th European Conference on Computer Supported Cooperative Work,
7-11 September 2009, Vienna, Austria


# Collaboration in Metagenomics: Sequence Databases and the Organization of Scientific Work

Matthew J. Bietz & Charlotte P. Lee
University of Washington
*{mbietz, cplee}@u.washington.edu*

**Abstract.** In this paper we conduct an ethnographic study of work to explore the interaction between scientific collaboration and computing technologies in the emerging science of metagenomics. In particular, we explore how databases serve to organize scientific collaboration. We find databases existing across scientific communities where scientists have different practices and priorities. We suggest while these databases appear to be boundary objects, they are better understood as boundary negotiating artifacts. Due to rapid scientific and technical innovation the tools, practices, and scientific questions change over the course of merely a few years resulting in challenges for collaboration.

## Introduction

The use of databases is critical for metagenomic science. While databases are often intended to span the boundaries between communities of practice (Wenger, 1998), they actually serve more as sites for the negotiation of scientific methods, research questions, and worldviews. Due to rapid scientific and technical innovation the tools, practices, and scientific questions change over the course of merely a few years. We find that multiple databases are useful for supporting work in a highly dynamic context of leading edge science. It is the production and use of these databases, and the implications therein for collaborative work and technology design, that will concern us in this paper.

Unlike traditional genomics, which focuses on the genetic code of specific organisms or species, metagenomics focuses on the distribution of genetic material within a population of microorganisms. The move toward metagenomic approaches has been enabled by technological advances in DNA sequencing that have allowed the generation of large amounts of data at significantly lower cost. These new technologies have created a number of cyberinfrastructure-related challenges, including exponentially increasing computational power and data storage needs and new algorithms for manipulating and analyzing data. Metagenomics also requires an interdisciplinary approach, frequently bringing together ecologists, geneticists, bioinformaticists, and computer scientists.

Our research suggests that metagenomics researchers, bioinformaticists, and developers of cyberinfrastructure often have a strong sense of an ideal conceptual Database that would contain all genetic sequence data and associated metadata, and often derivative analyses. This conceptual Database and implemented database systems frequently serve as a boundary negotiating artifacts (Bowker & Star, 1999; Lee, 2007). The multiplicity of databases, in particular, play a useful role in supporting highly innovative and dynamically changing activities.

# Background: Cyberinfrastructure and Databases

Cyberinfrastructures are distributed organizations supported by advanced technological infrastructures such as supercomputers and high-speed networks. Within the domain of scientific cyberinfrastructures (also known as e-Science), the capabilities of cyberinfrastructure are usually intended to be transformative (Atkins, et al., 2003). Cyberinfrastructures are employed to support work, often in the form of collaborative data sharing and, less frequently, analysis. The ability to pool data can enable scientists to answer questions that no single investigator or laboratory could answer individually. Large-scale data sharing can enable not only new types of scientific practices, but can also enable the exploration of new types of scientific questions. Conducting new types of science requires new and more powerful technologies to support communication, data sharing and analysis, and access to remote instruments.

The sharing of data, however, is rarely straightforward, Previous research has shown that that the development of effective CSCW systems to support data sharing groups requires a better understanding of the use of data in practice. Data play two general roles in scientific communities: 1) they serve as evidence to support scientific inquiry, and 2) they make a social contribution to the establishment and maintenance of communities of practice (Birnholtz & Bietz, 2003). Birnholtz and Bietz found that data sharing, particularly in fields with high task uncertainty, is problematic because of the difficulty of communicating contextual information in the absence of interpersonal interaction. Needed

contextual information includes the nature of the data, the scientific purpose of its collection, and the social function in the community that created it.

Issues of data sharing are critical to the development of large scale information infrastructures, but a treatment of data sharing should also engage a discussion of databases. In her work on databases as scientific instruments, Hine (2006) found a mouse genome database to be an emergent structure that is necessarily embedded in particular sets of work practices. She notes that:

> The patterns of connection and collaboration in scientific knowledge production involving databases can thus become both spatially and socially complex, building on existing networks but adding additional density, bandwidth and new tensions (Hine, 2006, p. 293).

In other words, the database is both built upon and enabling of scientific collaboration. The database provides both opportunities and constraints.

## Boundary Negotiating Artifacts

CSCW has long studied coordinative artifacts for the purposes of theorizing collaboration as well as informing the design of groupware. Many types of artifacts have coordinative functions and databases (not just the information contained therein) may be included among these. Research on coordinative artifacts have focused on paper and electronic documents (Lutters & Ackerman, 2002; Schmidt & Simone, 1996; Schmidt & Wagner, 2002) and have looked at these documents as boundary objects (Bowker, et al., 1999; Star, 1987-1989), boundary negotiating artifacts (Lee, 2007), and have put forth useful methodologies with which to understand how documents such as a report can serve to organize work in the most complex of organizations (Harper, 1998; Schmidt & Wagner, 2005). Many of these papers include in their purview spreadsheet documents and databased information (Harper, et al., 2001; Lee, 2007). Other fields have emphasized that the database itself is an important cultural form that entails a different mode of thinking about the world (Manovich, 2001) and as occasioning a new set of arrangements, as opposed to scientific journals for example, for the communication of scientific information and methods (Hilgartner, 1995).

Previous research has defined a *shared information system*, such as a shared database, to be an information system that is used by multiple communities of practice (Pawlowski, et al., 2000). These systems are described as typically spanning formal organizational boundaries such as functional departments or business units. Pawlowski et. al (2000) focus on enterprise-wide databases that support beginning-to-end business processes. They note that maintaining a shared system in an organization is challenging because triggers for system change can originate in any of the stakeholder areas when work practices or requirements change. Ultimately they suggest that shared information databases and related artifacts are boundary objects that require brokering, translating, coordinating and aligning perspectives, and addressing conflicting systems. Although we agree with

the larger premise that databases that are used by multiple communities of practice are key for boundary work and that these databases require brokering, a careful reading of the original work on boundary objects (Star, 1987-1989; Star & Griesemer, 1989) suggest that these databases may actually be a combination of boundary negotiating artifacts and boundary objects, or they may simply be boundary negotiating artifacts.

Defining features of boundary objects include that they pass from one community of practice to another with little or no explanation and satisfy the informational requirements of multiple communities of practice. Yet some of the things we call boundary objects do not actually do so (Lee, 2007). Throughout the literature described above, the following themes recur: So-called boundary objects may require considerable additional explanation and discussion to be intelligible; Artifacts sometimes play a role in the *active negotiation* of shared understanding amongst communities of practice (and thus can be used to enlist participation and can be adjusted through group interaction); Unstandardized artifacts that are partial, incomplete, or are intermediary representations are ubiquitous in collaborative work; And so-called boundary objects can "fail" to satisfy the informational needs of collaborating parties (Henderson, 1999; Lee, 2007; Subrahmanian, et al., 2003). The recurring themes described here indicate that the boundary objects concept is not incorrect, rather it is incomplete. Other researchers have grappled with fitting their research findings to the notion of boundary objects. Henderson (1999) found that the boundary object concept required amendment and suggested the term *conscription devices* to refer to a type of boundary object that enlists group participation and that are adjusted through group interaction. Subrahmanian et al. (2003) proposed the broad concept of *prototypes* based on their observation of artifacts that support systematic updating of boundary objects. Organizational changes, they note, sometimes caused boundary objects to be unable to support activity. O'Day et al. (2001), in their work on molecular biologists and computer scientists, refer to *boundary objects in-the-making* which are unstable objects that still work to facilitate collaboration across communities by giving people common ground for discussion and negotiation. They note that in the absence of durable cooperation, boundary objects in-the-making are necessary to confront and reconcile different local meanings.

We stress the importance of adopting a strict definition of boundary objects that is true its origins. By doing so, we can fully appreciate just how large and nuanced is the research and design space when we accept the idea that many artifacts and practices are not just crossing but weaving, pushing, pulling, and everything else on, around, and through communities of practice. Boundary negotiating artifacts provide a lens through which we can view the myriad uses of artifacts, many of them messy and ad hoc. Boundary negotiating artifacts:

- Are surrounded by sets of practices that may or may not be agreed upon by participants

- Facilitate the crossing of boundaries (transmitting information)
- Facilitate the pushing and establishing of boundaries (dividing labor)
- May seem "effortful" in use as opposed to effortless
- Are fluid: 1) a boundary negotiating artifact can change from one type to another when the context of use changes; and 2) a boundary negotiating artifact can sometimes also simultaneously be physically incorporated or transformed into another artifact
- Can be largely sufficient for collaboration
- Are possible predecessors of boundary objects (Lee, 2007)

Boundary negotiating artifacts are used to: record, organize, explore and share ideas; introduce concepts and techniques; create alliances; create a venue for the exchange of information; augment brokering activities; and create shared understanding about specific problems. Scientific collaboration between biologists and computational disciplines have been noted as an endeavor that requires interpretive frames to be brought together:

> At the end of the day, people in biological and computational disciplines try to produce biological understanding by bringing their distinctive interpretive frames together. But as we have discussed, it is likely that there will be an ongoing need for negotiation between disciplines. It is not the case that biologists can simply learn how to run the numbers; the numbers and way to run them continue to be problematic as biologists ask new questions and encounter new forms of data. Similarly, mathematicians and computer scientists are challenged to develop new analytical methods to deal with the flexibility and multi-dimensionality of living systems. Biologists and computation experts need to continue their collaboration. (O'Day, et al., 2001, p. 417)

In the research that follows, we too find that the practices and interpretive frames of biologists and computational disciplines need to be brought together, but in the area of metagenomics research there is even greater collaborative complexity. We find that the communities of biologists who are attempting to share databases have diverse practices, interpretive frames, and different scientific concerns that are brought together by databases which function as boundary negotiating artifacts.

# The Current Study

We employed ethnographic research methods which involved entering into sites involved in the production of metagenomics research and databases, getting to know the people involved, participating in the daily routines of the setting, and observing what is 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 engagement with these communities began in the summer of 2007, and is ongoing at the time of writing. Our initial focus was on one particular metagenomics database project, and our primary focus for the first year of our engagement was on the team developing the database. We interviewed as many members of the development team as we could, some of them multiple times. For four months of this time, one of the authors attended weekly project meetings, ad

hoc meetings, and spent at least one day per week working from an assigned desk in the development team area.

In the second year of engagement, our focus shifted to developing a broader understanding of the landscape of cyberinfrastructure for metagenomics research. We interviewed microbiologists, bioinformaticists, computer scientists, and representatives of funding agencies. We interviewed both users and developers of several major genomics and metagenomics databases. We attended conferences and workshops devoted to metagenomics research, database development, and the development of standards. For the past seven months, one of the authors has attended weekly laboratory meetings at an academic molecular biology laboratory engaged in metagenomics research.

In total, this amounts to thirty-three formal interviews and well over one-hundred hours of on-site observation and informal conversation. Interviews were semi-structured and ranged from thirty minutes to nearly two hours, with most lasting between sixty and ninety minutes. Transcriptions of the interviews, field notes, and various indigenous documents were coded in Atlas.ti using a grounded theory approach (Glaser & Strauss, 1967).

# A Metagenomics Primer

The term "metagenomics" was coined in 1998, and while there is some controversy about the exact definition of the term, it generally refers to using genomics techniques to study communities of microorganisms (Chen & Pachter, 2005; Handelsman, et al., 1998). Until recently, it was necessary to culture microorganisms in a laboratory in order to produce enough DNA for sequencing. However, it is estimated that less than one percent of the world's microorganisms can be cultured using standard laboratory techniques (Hugenholtz, et al., 1998). Advances in DNA amplification techniques and new sequencing technologies have significantly reduced the cost of sequencing and made it possible to analyze DNA without culturing, giving scientists access to a newfound wealth of genetic information.

Metagenomic techniques are relevant to a number of fields, including marine ecology, medicine, energy production, and environmental remediation, to name a few. In a typical metagenomic experiment, scientists begin by sampling the microorganisms from a particular environment. For example, a marine microbiologist may pass seawater through a series of progressively smaller filters to isolate a particular kind of microorganism (viruses, bacteria, etc.). DNA is then extracted from the organisms and prepared for sequencing. While the specifics vary across manufacturers and technologies, many metagenomic analyses use "shotgun sequencing," in which the long strands of DNA are randomly broken up into shorter segments which are "read" by the sequencer. Depending on the technology, these "reads" range from 20 to 400 base pairs in length. Longer

segments of DNA are then computationally reconstructed by searching for areas of overlap among the shorter segments.

Because of the number of different organisms in the sample, this technique typically results in only a small portion of each organism's genome being sequenced. Through a combination of statistical techniques and comparisons to known genomes, scientists can identify the most prevalent organisms in their sample and estimate the diversity of organisms in the population. Scientists can also study the functional capacities of the population of microorganisms in relation to their environment, for example, the ability of marine microbes to metabolize phosphorous (Gilbert, et al., 2009) or the influence of the microbial population of the gut on obesity (Turnbaugh, et al., 2008).

## Computation in Metagenomics

A discussion of metagenomics practice would be incomplete without a discussion of the computational resources on which metagenomics relies. Metagenomics would not be possible without a broad array of computational tools and information systems. Computation is so central to the work of these scientists that most of our senior biologists spent very little time at "the bench" working with wet materials. Although their students spent more time at the bench and in the field collecting samples, our senior participants all reported spending at least 90% of their research time at a computer.

Our participants often resort to the metaphor of jigsaw puzzles to explain the role of computation in metagenomics. Environmental shotgun sequencing has been compared to mixing the pieces from many different jigsaw puzzles in the same bag and pulling out a few handfuls of the pieces. The computer is used to put the pieces together when possible, and from the resulting fragments of puzzles, try to figure out how many puzzles were in the bag and what picture was on each one.

So the first computational task is to assemble the fragmented DNA sequences (the puzzle pieces) into longer contiguous sequences. This is made more complicated by differences among sequencing technologies, which result in varying read lengths and error rates. Even so, assembly is seen to be a relatively straightforward process compared to the later analysis of the assembled sequences.

Frequently the next step is to make the assembled sequences biologically meaningful by "annotating" them. During annotation the sequences are analyzed and compared to existing sequence data in order to identify regions of the genetic code that we already know something about. Depending on the tools used and the scientific goals, annotation may identify the physical structure of the DNA, its functional properties (e.g. what proteins it produces), or even the organisms that are known to have this particular sequence. Annotation is both a computation- and data-intensive process. Successfully annotating sequences requires comprehensive and well-curated database of known sequences to which the new sequence can be

compared. And even the most powerful automated annotation systems require several hours to several days to annotate the data produced in a single run on a current DNA sequencing machine.

After annotation, researchers will analyze the annotated sequences using statistical analysis packages and visualization tools. While there are some "off-the-shelf" packages available, frequently these analyses are conducted using custom software and analysis scripts.

For all of these steps, but especially for annotation and certain forms of statistical analysis, the scientist must compare the sequences they are studying to other known sequence data. The need to assemble, collect, compare, and annotate large volumes of DNA sequence data precipitates numerous databases. How these databases are produced and used have implications for collaborative work and technology design.

# The Ideal Database

An underlying theme across our participant interviews is the notion of an ideal Database.[1] Our participants talked about being able to share data, across what we in CSCW would describe as communities of practice, implying that existing databases are serving as boundary objects and consequently satisfying the informational requirements of all. Further investigation, however, shows that the notions of the Database are highly idealized and when delving into the details of practice, the successful use of these databases requires a great deal of translational and interpretive work.

Scientists using metagenomic approaches have a strong sense that sequence data is a public good.[2] The scientists we interviewed are keenly aware that to conduct research in this area it is necessary to have access to the data of others in order to compare genes at hand against previously found genes. In order to gauge environmental trends across time and space and to ascertain the unique qualities of particular genes requires access to amounts of data so vast that no single researcher or group of researchers could collect enough data. The "Database" is a particularly evocative concept here: a key feature of metagenomics research is that all prior sequence data serves as the baseline against which new sequences are compared.

A biologist working on the design of a database system described one scientific rationale for creating collections of sequence data:

---

1 We use a capital "D" when we are referring to the conceptual ideal Database, and a small "d" when referring to a specific database system.

2 It should be noted that most of our informants were working on government- or foundation-sponsored research in academic settings. However, even among those scientists who were involved with commercial research, we found that proprietary concerns might delay, but usually would not prevent the public release of sequence data.

> In order to understand a new gene that we don't know what it does, we need to compare it with all the other genes that we have in the database that we know what they do. So we know what to do, for example, because they have been experimentally verified. We compare the sequences, computationally and we find the sequence in the database, and based on that, we can predict what the function of the gene may be. So then, what we also do is we integrate all that data in a single database because that is what is facilitating the comparative analysis; you must integrate all the data.

Comparing a new sequence against the Database can reveal the identity and function of genes and organisms. Similar comparisons to and analyses of "all the data" are used to understand the evolutionary history of organisms or the diversity of microbial populations. These comparisons are even useful for determining if a gene has been previously identified for patent applications.

One of the themes that emerged from our interviews and observations is that metagenomics researchers, bioinformaticists, and developers of cyberinfrastructure often have a strong sense of an ideal Database that they would like to have available or are actively trying to create. While the details of the ideal Database vary from person to person, generally it holds all available sequence data and associated metadata, and often derivative analyses. The Data would be well classified and annotated, and the Database would not contain errors or redundancies.

The ideal Database is also explicitly collaborative. The Data would be collected from and useful to scientists from a wide variety of communities of practice. Part of the rationale for spending the large sums of money required to develop such comprehensive databases is because the Data could reach across so many domains. The same sequence Data are potentially useful for medical research, environmental remediation, energy production, national security, drug development, chemical production, and many other pursuits. The Database is intended to be a boundary object, providing a standardized repository supporting cooperation across multiple communities of practice.

Our informants tend to think of the ideal Database as separate from the specific database systems they use in their work. One metagenomics researcher spoke of the Database this way:

> We also rely on data that is in the database…. We generate a lot of primary data ourselves, but if we want to make comparisons, we have to compare to what's in the database. So we will use EMBL and GENBANK and the data that's in those databases as well.

This scientist refers to all of the sequence data produced outside of his laboratory as "*in the database*." He then goes on to list specific sequence database systems he uses. This was a common trope across many of our interviews with scientists. When we asked them to describe the process of analyzing sequence data, they would often say that they compared the sequence data they generated in their laboratories to "all the other sequence data" or "every other known sequence." On the other hand, when we observed their work or asked the scientists to tell us

about the specific databases they used, we found that they often used multiple databases, none of which actually contained all the Data.

The lack of integration of databases and datasets creates usability problems for scientists. As no single database contains all of the Data, scientists will often create their own local aggregated datasets to work with:

> It's very important to have all metagenomes gathered together in one platform so that when people look for metagenomes they don't have to go here and here and, you know, it makes it all convenient… Yes, if you were doing a complex analysis and gathering data from many metagenomes and you would have to register on this server and this server and also this server to get the metagenomes. And then this server would provide you with some information, this other one slightly different information and the third one another kind of different information. It just makes it really hard if the data is all scattered around.

There are a number of projects that are working to make the ideal sequence Database a reality. Our respondents reported using many other database systems including GENBANK, the EUROPEAN MOLECULAR BIOLOGY LABORATORY NUCLEOTIDE SEQUENCE DATABASE (EMBL-BANK), the COMMUNITY CYBERINFRASTRUCTURE FOR ADVANCED MARINE MICROBIAL ECOLOGY RESEARCH AND ANALYSIS (CAMERA), THE SEED, INTEGRATED MICROBIAL GENOMES (IMG/M) and others. Probably the most well-known and longest-lived example is GENBANK, which was created in 1982, and is now housed at the National Center for Biotechnology Information, a division of the National Library of Medicine at the National Institutes of Health. GENBANK was founded to be "an annotated collection of all publicly available DNA sequences" (National Center for Biotechnology Information). A developer of another database system told us about the amount of effort spent to try to create a comprehensive database:

> We're very aggressive in going out and getting basically all the data that relate to the public domain and integrate them. This is one of the most intense parts of maintaining and updating the system, constantly updating and adding everything that is released to the public domain.

This is made more difficult because sequence data are being generated in many locations, and even GENBANK does not contain all of the publicly available sequences:

> There are several other sequencing centers that do not directly submit their data into GENBANK. They are keeping the data and releasing them through their websites, but they are not necessarily depositing them directly into GENBANK.

But even with the incompleteness of individual database instantiations, participants still expressed confidence in the ideal Database. One developer of a competing database systems told us:

> It doesn't serve the community well if [our database] stands out there distinguished, beating its chest, saying we have more data than [GENBANK] or we have different data than [GENBANK]. I would argue philosophically that's a losing strategy and [our database] should not distinguish itself on what data it contains.

Many of our informants felt that what "the science" and "the community" required was for all of the specific sequence database systems to be operating on the same set of Data. Projects are underway to facilitate the creation of this universal

Dataset across database systems. For example, the GENOMIC ROSETTA STONE "is creating a mapping of identifiers describing complete genomes across a wide range of relevant databases so that information about genomes and the organism from which they derive can be more easily integrated" (Genomic Standards Consortium, 2008). Their vision is to create a distributed but easily accessible version of the universal Dataset by connecting many database systems into a federated database.

However committed scientists and database developers are to realizing a concrete version of this abstract ideal Database, we find that the vision for the Database is contested. Both the ideal Database and particular database systems are implicated in ongoing controversies about appropriate research questions, the role of the researcher, science funding, and scientific validity. At the same time that the Database supports collaboration, it is also playing a role in the active negotiation of practices and understandings. Rather than passing easily between communities of practice, using the databases requires significant translational work. Every scientist we spoke with reported using multiple databases, often having to manually reformat, edit, and combine the outputs of different databases. The databases often do not contain the contextual information necessary to make sense of the sequence data. Frequently this results in frustration for both users and developers.

> The big issues with metagenomics is that the big archives are dysfunctional. They're not only dysfunctional for metagenomics, they're also dysfunctional for genomics these days.

The Database is intended to be a boundary object, but we believe that it is more productive to understand both the ideal Database and the individual instantiations as boundary negotiating artifacts.

# The Database as Boundary Negotiating Artifact

Participants describe the individual sequence database systems as if they were shadows, poor representations of a widely-agreed-upon ideal. We find, however, that by looking across the landscape of databases, a different picture emerges. Instead, each decision about the implementation of a particular database system plants a stake for a community boundary. The databases are not so much imperfect copies of an ideal as they are arguments about what the ideal Database should be.

In this section, we will draw on our observations and interviews to discuss two areas of negotiation around the Database. First, we will discuss the close relationship between the Data and local scientific practice. Then we will discuss the problem of metadata and information completeness in sequence databases. In both instances, we find that rather than being stable boundary objects that move across community boundaries, the databases are malleable artifacts that serve as sites for negotiation of community boundaries.

## The Database in Practice

The Database both contributes to and results from scientific practice. To understand this claim, it is important to look a little deeper at the technical implementation of what are commonly called databases. Our respondents speak of the databases as collections of data, but that only tells part of the story. It is more accurate to think of them as database-driven systems that include some combination of raw sequence data, contextual metadata (data about the environment from which the sample was collected), procedural metadata (data about how the samples were processed), assembled sequences, annotations, pre-computed analyses, and various tools for data comparison, annotation, visualization and analysis.

The particular arrangement of data and tools that make up each of these systems is driven by particular scientific needs. One scientist involved in database development told us:

> So I kind of think it goes back to having that question, right…. What's your underlying emphasis for having the database? So, our underlying emphasis is that we have some questions that we're trying to answer both in complete genomes and for metagenomes…. Some of the things that we're trying to do is to take really specific problems that we're trying to address and use [our database system] to address some of those problems.

One of the ways that databases are tuned for particular research questions is through their accession policy. A typical strategy is for databases to focus on a particular type of organism or environment. For example, there are databases that focus on marine microorganisms, soil microorganisms, organisms found in the human gut, etc. Another database is attempting to collect data only about pathogenic organisms.

Another strategy is to focus on a particular type of data, regardless of the source. For example, some databases are collecting only "16S ribosomal RNA" sequences. These sequences are subunits of RNA that are useful for studying the evolutionary relatedness of species. But these sequences (which also appear in more general archives like GENBANK) are applicable only to specific kinds of research questions. One scientist who studies microbes that cause various diseases explained why she did not use 16S databases:

> Everybody uses 16S and 18S sequences to categorize the phylogenetic community present. But similar organisms may have the same 16S, but have completely different physiology. So some, like *vibrio* for example, they're a great example of this. Many *vibrios* have the same 16S but can acquire a few genes, either by horizontal gene transfer, by phage transfer and they become highly virulent. *Vibrio cholerae* is a great example of that. You can have *vibrio cholerae* that's not toxic at all. It acquires one gene from its phage, the CTX gene. Horribly virulent organism, but if you look at the 16S, you'll never know.

In other words, the method used by many metagenomics researchers to categorize an organism, is useless for certain types of questions such as those about whether or not an organism is toxic or infectious. 16S sequence databases are useful for understanding how species relate to each other, but they are not sufficient for

understanding how variations in other parts of the genome can lead to functional differences in microorganisms. By choosing to only collect 16S sequences, the database developers have privileged certain scientific questions over others.

Database systems are also customized with particular query and analysis tools. One advisor to a database system told us about the problem of inheriting data and tools from a different research community:

> So they had a lot of approaches to data analysis. Now, what they were looking for was slightly different than what the [microbial ecologists] were looking for, but it was something that many people were interested in. They were basically on a hunt for genes…. Eventually, all of the data [they] had in hand plus all of their analysis, eventually became the first datasets in [our] database. And many of the database structures in the database tools were developed by [them]. And for the purposes that they had at the time they did that, I think the database was actually adequate and not too bad. The problem was that it didn't serve [our] community quite as effectively as one would like. And so we made a series of recommendations over time about restructuring the system to be more accommodating to the kinds of questions that the ecologists were asking rather than the kinds of questions that molecular biologists and gene finders were asking.

When the microbial ecology project adopted the database system from the traditional genomic "gene finders," they expected the database to be a boundary object. They knew they would have to customize it to some extent, but thought it would be able to "travel across borders and maintain some sort of constant identity" (Bowker, et al., 1999, p. 16). In the end, however, the system was so tailored to a specific set of research questions that the collection of data, the set of tools, and even the social organization of the project had to be significantly changed. New analysis tools were developed and old tools were discarded. Not only was the database ported to a different technology, the data itself was significantly restructured to fit the new tools and approaches. While the database development projects had begun by working together, in the end they were unable to collaborate. The system that was supposed to tie these groups together could not be shielded from the controversies that formed the boundaries between the communities of practice.

## Metadata and Informational Needs

One of the features distinguishing metagenomic approaches from traditional genomics is a reliance on contextual data, or *metadata*. Unlike traditional genomics that focuses on the genetic information in a single organism, metagenomics considers the relationships of populations of microorganisms to their environments. In order to understand, for example, the effect of changing ocean temperatures on microbial populations, it is necessary not only to have sequence data but also to have associated data about where and when the sample was taken. Ideally, every sequence in the database would be linked to data about the environment from which the sample was taken, the people involved in the samples collection and processing, and the procedures used to isolate and

sequence the DNA. But collecting, storing, and disseminating this metadata adds another layer of complexity to the technical exercise of database development. Metadata and metadata standards become contested artifacts and sites of negotiation within the metagenomics and wider genomics communities.

One of the defining characteristics of boundary objects is that they are able to satisfy the information needs of different communities of practice. However, changes in the information needs of the community and the inclusion of new communities can challenge the ability of an artifact to be a boundary object. Metagenomics brings new questions, and existing sequence databases are inadequate for the metagenomics community's needs.

Until recently, most existing sequence databases had little, if any, metadata support. Even if scientists wanted to share metadata through the database, often their only recourse was to add a comment in a free-text field. More commonly, a scientist wanting to know more about a sequence in the database would have to track down associated publications and hope that the authors had included the relevant details. A program officer from a funding agency described the message coming from metagenomics researchers:

> 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."

Not only were there no metadata-capable databases, but there were no standards for what contextual data should be collected or how to represent it for storage. Scientists will typically only collect the data that is relevant for the study at hand. One scientist expressed frustration about the difficulty of sharing metadata:

> You don't measure salinity when you work in the ocean. Right? You just assume the salinity is about the same…. Unless you've got a CTD [conductivity, temperature, and depth sensor] or something…. It really depends on what your question is. What I think is important as metadata, in fact what I know is important as metadata, nobody will ever measure…. We're doing microbial ecology. Essentially nobody measures what the microbes are eating…. That's because it's a hard thing to measure. But they'll all have nutrient analysis, though. That's because nutrient analyses are easy to measure.

Having metadata standards is important for both scientists and database developers. For scientists, a standard can function as a guide for what data to collect and how to represent it. For the database developers, the standard outlines what data should be in the database. Metadata standards are in active development, and some have even been published (Field, et al., 2008), but these standards are still being negotiated and none have been widely adopted.

But the adoption of these standards reveals the way that the Database not only crosses boundaries but is also implicated in pushing and establishing boundaries. Environmental metadata is extremely important for microbial ecologists, is less important for some other metagenomic questions, and is significantly less

important to many traditional gene- and whole-genome-focused users of the Database. The upshot is that it is important to the microbial ecologist that the geneticist attaches environmental metadata to sequence data, but it is not important to the geneticist. Similarly, the metadata needed by a marine microbiologist is significantly different from that needed by someone studying the microbial population of the human gut. This is a classic case of a disparity between those who must do extra work and those who benefit from the work (Grudin, 1989).

In the face of this difference in the value of metadata between communities of practice, the database becomes an important site for negotiations of the division of labor. In a discussion of metadata standards, developers of sequence databases were asked to require contributors to submit standards-compliant metadata with their sequence data. Databases that could not (or would not) make metadata a requirement were asked to alter the interface to make metadata submission easier, to make it easier to limit searches to sequences with metadata, and to create certification programs to give special status to sequences with compliant metadata. To use the language of Latour (Latour, 1987), the database becomes a mechanism for enrolling and controlling others in the creation of a particular kind of science.

# Supporting Collaboration

While biologists, computational biologists, bioinformaticists, and computer scientists take the need to work together as given, collaborative endeavors differ according to content and scope. What matters are the particular scientific questions, not disciplinary allegiances or training. Each set of scientific concerns requires different types of metadata and different types of output. The database is a common denominator but is not sufficient for accomplishing work. Collaboration in the metagenomics area can be crudely classed according to whether they prioritize biological or environmental questions, but upon further investigation those classes quickly breakdown into subcategories with some overlapping and some unique requirements.

We found that the fit between the database system and the scientists' research questions was a more important decision factor in choosing a database to use than the completeness of the database. One researcher told us about certain databases being better repositories than others, but then when asked why he chose to use particular databases, he said:

> Researcher: Because of what it does. Because of what I can get out of it.
>
> Interviewer: Is it about the tools or the data that they have?
>
> Researcher: It's about the results. The different websites - you can get the data from any of those websites. It's about the tools. It's about the results that they can produce for you.

Another researcher also emphasized the importance of the visualization and analysis tools:

> Well, originally I started using [that] database because it's a great way to look at functional analysis…. What you're looking at is at the functional profiles of each one of these samples. And like I said before, that's really important for understanding the function of a community…. And so we can track - and perhaps if you think of an ecosystem, or in this case, a metabolic system, looking at shifts in the metabolism of the whole system might be more environmentally relevant than just looking at the change in a particular strain of bacteria. And so you can see these massive changes; all sorts of great ways to parse the data into something that's biologically relevant.

For this scientist, the most important criteria for choosing this database was the ability to analyze and visualize the data in a way that made it "biologically relevant" to the questions she wanted to answer. For these scientists, the best database was not the one that came closest to the ideal comprehensive "all known sequences" Database. Instead, it was the one that best fit the research, in other words, the one where the entire database system—data, structures, tools, and outputs—came together to best support the scientist's practice and produce the most meaningful answers to scientific questions.

Designing databases that work for scientists is an immense challenge. There is a great diversity of need regarding data, metadata, and software tools that stems from a diversity of scientists and scientific interests. Advances have been made with technologies like ontologies, which can provide semantic mappings across domains (Schuurman & Leszczynski, 2008). But the challenges described here go beyond semantics and invoke questions of the value and organization of scientific practice. A metagenomicist who has collaborated in the development of a database system told us about three stakeholder communities that are trying to use a particular database. He describes the groups as moving targets:

> There are at least three moving targets in this project. And that is that there are the ecologist metagenomics people, there are evolution people that are more interested in the evolution of the sequences, you know, what they're telling you about evolution; which is actually quite different how you analyze the data in this case. And then there are just the people that are thinking, like, just genomes and glorified genomes, right. And that's also a very different way of looking at the world. I think that that's a big failing that we didn't recognize that in the beginning as much as we should have….
>
> We need different outputs. That's kind of the problem. So you do almost the same thing in the beginning, but if you're interested in the genome, you want a genome browser, right, you want to scroll on a genome and look at a gene, where it's at and everything. If you're someone like me, you want something that can be funneled into a statistical package. And if you're an evolutionist, you want the same information, but you want to be able to do an alignment with them. I mean it's the same exact analysis, but a different 'what do you do with it at the end' sort of thing…. It's the tools that really count.

What we see is that there are a number of communities using these database systems, and each brings its own set of research questions and viewpoints. At a base level they are all using the "same" sequence data. But in practice, the Data do not exist independently of the database system. Even accessing "raw" data

requires understanding the particular data collection in the given system (along with its potential errors, omissions, and redundancies), navigating a particular set of data standards and formats, and dealing with the particular query and output technologies. When a system lets you select data based on the presence of a particular gene but not based on the geographic location from which they were sampled, that system reinforces a particular set of research questions and strengthens the boundaries between communities of practice.

While there is a diversity of scientists and scientific interests, the larger challenge is that the research questions are continually evolving.

> The instruments will continue to improve. But they're never going to be perfect because we're continuing to push the boundaries. So the kinds of scientific questions we can answer will keep extending. So we'll have demands for new instrumentation. We'll have demands for new software tools. And I can make up 20 questions that are important today. Any microecologist can make up better questions, probably, than I can about really - some the same, some different and better than I can about the challenges; the new questions that metagenomics will allow us to ask and answer. But I think we also feel that we don't know the range of questions fully. And so the same is true for software tools.

As scientists are successful at generating new discoveries, as technologists are successful at developing new technologies, and as these innovations synergistically drive each other forward, the research questions will change and the range of answerable research questions will also change. Many of our participants discuss being drawn to "metagenomics tools" and some of our participants refer to metagenomics as a new discipline. It is not a far stretch then to assume that as research questions, data, tools, and practices shift and change so too will the communities around them.

Those wishing to support scientific collaboration should take care to map out scientific stakeholders according to scientific questions, and not according to domain or institutional allegiances. This mapping out of concerns must be done iteratively to keep pace with scientific developments. Furthermore, CSCW researchers in this area should be aware that scientists will talk about databases as if they function as boundary objects, but that when pressed for more detail, scientists reveal that their databases require a great deal of work in order to meet the needs of different communities of practice. This latter phenomenon is not necessarily a failure of requirements specification. The requirements are often sufficient at the time of collection but are rendered inadequate by scientific advances. Furthermore the multiplicity of databases is a reflection of the multiplicity of interests and competing knowledge claims that are indicative of a vigorous scientific community. Some degree of integration may be desirable, even inevitable, but a high degree of integration among research databases is a mirage—a utopian ideal. A lofty goal for computer supported cooperative science would be to find ways to support simultaneously cooperative data sharing and scientifically competitive (i.e. divergent or unique) data acquisition use, analysis, and theory building.

# Conclusion

New types of science also require new standards, processes, and collaborative social structures, such as distributed virtual organizations comprised of domain scientists, information scientists, and engineers. Metagenomic science is among many endeavors that require work to be coordinated through and around multiple databases. More work is needed to understand how collaborative work is structured by multiple databases. The tendency to dismiss situations where organizations depend on imperfectly interoperable databases as merely inefficient legacy systems is likely glossing over insights about just how multiple databases support not only different types of work but also different perspectives and priorities. There is a more interesting story to tell about how they actually serve to support and constrain work. We also find that there is an important connection between multiple databases and coordinative artifacts.

As mentioned in earlier research on boundary negotiating artifacts in a small, nascent design group using primarily paper documents, artifacts can be used to cross boundaries between communities of practice. But they can also be used to affect the division of labor, or in other words, to push and establish the boundaries between communities of practice (Lee, 2007). In this paper we have looked at metagenomic science, which is a very different sort of endeavor, and yet it too requires complex coordination around another type of artifact: the database. Rather than looking at databases as static, we choose to look at databases as existing across scientific communities where the scientists involved have different practices and priorities. Due to rapid scientific and technical innovation the tools, practices, and scientific questions change over the course of merely a few years. The sequences within the databases are relatively static, being a sort of minimum common denominator, but what make the databases useful and relevant are the array of constantly-changing software tools and highly negotiated metadata.

In dynamic environments, the number of true boundary objects that satisfy the information requirements of multiple communities of practice may be relatively few compared to the number of prototypes, boundary objects in-the-making, or boundary negotiating artifacts. If we can consider the database to be another type of artifact that coordinates multiple perspectives, we begin to see how multiple databases may sometimes be necessary and useful. The challenge for computer supported cooperative science then becomes how to meaningfully support large-scale collaborations that are reliant on multiple databases that support a multiplicity of knowledge building priorities and practices.

# Acknowledgments

Special thanks to the participants who were so generous with their time, Eric Baumer, who assisted in data collection, and the anonymous reviewers for their comments. This work was supported by National Science Foundation awards IIS-0712994 and OCI-083860.

# References

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messina, P., 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.

Birnholtz, J., & Bietz, M. J. (2003): 'Data at work: Supporting sharing in science and engineering' *Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work*, New York, NY: ACM Press, pp. 339-348.

Bowker, G. C., & Star, S. L. (1999): *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

Chen, K., & Pachter, L. (2005): 'Bioinformatics for whole-genome shotgun sequencing of microbial communities', *PLoS Computational Biology,* vol. 1, no. 2, Jul, pp. 106-112.

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,* vol. 26, no. 5, May 2008, pp. 541-547.

Genomic Standards Consortium (2008): 'Genomic Rosetta Stone', Retrieved March 5, 2009, from http://gensc.org/gc_wiki/index.php/Genomic_Rosetta_Stone

Gilbert, J. A., Thomas, S., Cooley, N. A., Kulakova, A., Field, D., Booth, T., et al. (2009): 'Potential for phosphonoacetate utilization by marine bacteria in temperate coastal waters', *Environmental Microbiology,* vol. 11, no. 1, Jan, pp. 111-125.

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

Grudin, J. (1989): 'Why groupware applications fail: problems in design and evaluation', *Office: Technology and People,* vol. 4, no. 3, pp. 245-264.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998): 'Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products', *Chemical Biology,* vol. 5, no. 10, Oct, pp. R245-249.

Harper, R. (1998): *Inside the IMF: An Ethnography of Documents, Technology and Organizational Action*. San Diego: Academic Press.

Harper, R., Procter, R., Randall, D., & Rouncefield (2001): "Safety in numbers': Calculation and document re-use in knowledge work' *Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work*, New York: ACM, pp. 242-251.

Henderson, K. (1999): *On Line and On Paper: Visual Representations, Visual Culture, and Computer Graphics in Design Engineering*. Cambridge, MA: MIT Press.

Hilgartner, S. (1995): 'Biomolecular databases: New communication regimes for biology?', *Science Communication,* vol. 17, no. 2, pp. 240-263.

Hine, C. (2006): 'Databases as scientific instruments and their role in the ordering of scientific work', *Social Studies of Science,* vol. 36, no. 2, April 1, 2006, pp. 269-298.

Hugenholtz, P., Goebel, B. M., & Pace, N. R. (1998): 'Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity', *Journal of Bacteriology,* vol. 180, no. 18, pp. 4765-4774.

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

Lee, C. P. (2007): 'Boundary negotiating artifacts: Unbinding the routine of boundary objects and embracing chaos in collaborative work', *Computer Supported Cooperative Work: The Journal of Collaborative Computing,* vol. 16, no. 3, pp. 307-339.

Lutters, W. G., & Ackerman, M. S. (2002): 'Achieving safety: A field study of boundary objects in aircraft technical support' *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, New York: ACM, pp. 266-275.

Manovich, L. (2001): *The Language of New Media*. Cambridge: MIT Press.

National Center for Biotechnology Information (April 2, 2008): 'GenBank Overview', Retrieved February 23, 2009, from http://www.ncbi.nlm.nih.gov/Genbank/index.html

O'Day, V., Adler, A., Kuchinsky, A., & Bouch, A. (2001): 'When worlds collide: Molecular biology as interdisciplinary collaboration', in W. Prinz, M. Jarke, Y. Rogers, K. Schmidt & V. Wulf (eds.), *Proceedings of the Seventh European Conference on Computer-Supported Cooperative Work*, Dordrecht, Netherlands: Kluwer, pp. 399-418.

Pawlowski, S. D., Robey, D., & Raven, A. (2000): 'Supporting shared information systems: Boundary objects, communities, and brokering' *Proceedings of the 21st International Conference on Information Systems*, Atlanta, GA: Association for Information Systems, pp. 329-338.

Schmidt, K., & Simone, C. (1996): 'Coordination mechanisms: Towards a conceptual foundation of CSCW systems design', *Computer Supported Cooperative Work (CSCW),* vol. 5, no. 2, pp. 155-200.

Schmidt, K., & Wagner, I. (2002): 'Coordinative artifacts in architectural practice', in M. Blay-Fornarino, A. M. Pinna-Dery, K. Schmidt & I. Wagner (eds.), *Cooperative Systems Design: A Challenge of the Mobility Age*, Amsterdam, The Netherlands: IOS Press, pp. 257-274.

Schmidt, K., & Wagner, I. (2005): 'Ordering systems: Coordinative practices and artifacts in architectural design and planning', *Computer Supported Cooperative Work: The Journal of Collaborative Computing,* vol. 13, no. 5-6, pp. 349-408.

Schuurman, N., & Leszczynski, A. (2008): 'Ontologies for bioinformatics', *Bioinformatics and Biology Insights,* vol. 2008, no. 2, pp. 187-200.

Star, S. L. (1987-1989): 'The structure of ill-structured solutions: Boundary objects and heterogeneous distributed problem solving', in L. Gasser & M. N. Huhns (eds.), *Distributed Artificial Intelligence*, San Mateo, CA: Morgan Kaufmann, Vol. II, pp. 37-54.

Star, S. L., & Griesemer, J. R. (1989): 'Institutional ecology, 'translations' and boundary objects: amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39', *Social Studies of Science,* vol. 19, no. 3, pp. 387-420.

Subrahmanian, E., Monarch, I., Konda, S., Granger, H., Milliken, R., Westerberg, A., et al. (2003): 'Boundary objects and prototypes at the interfaces of engineering design', *Computer Supported Cooperative Work: The Journal of Collaborative Computing,* vol. 12, no. 2, 2003, pp. 185-203.

Turnbaugh, P. J., Backhed, F., Fulton, L., & Gordon, J. I. (2008): 'Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome', *Cell Host Microbe,* vol. 3, no. 4, Apr 17, pp. 213-223.

Wenger, E. (1998): *Communities of Practice: Learning, Meaning, and Identity*. New York: Cambridge University Press.