Data at Work: Supporting Sharing in Science and Engineering

Jeremy P. Birnholtz
School of Information, University of Michigan
1075 Beal Ave., Ann Arbor, MI 48109
+1 734-647-8044
jbirnhol@umich.edu

Matthew J. Bietz
School of Information, University of Michigan
1075 Beal Ave., Ann Arbor, MI 48109
+1 734-763-2285
mbietz@umich.edu

ABSTRACT
Data are a fundamental component of science and engineering work, and the ability to share data is critical to the validation and progress of science. Data sharing and reuse in some fields, however, has proven to be a difficult problem. This paper argues that the development of effective CSCW systems to support data sharing in work groups requires a better understanding of the use of data in practice. Drawing on our work with three scientific disciplines, we show that 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. A clearer consideration and understanding of these roles can contribute to the design of more effective data sharing systems. We suggest that this can be achieved through supporting social interaction around data abstractions, reaching beyond current metadata models, and supporting the social roles of data.

Categories and Subject Descriptors
H.5.3 [**Group and Organization Interfaces**]: *Computer-supported cooperative work*

General Terms: Design, Human Factors, Theory.

Keywords: Data Sharing; Communities of Practice; Metadata; Collaboratories.

1. INTRODUCTION
Collaboration and social interaction are essential to modern scientific practice [26, 28]. Traditionally, this collaboration has occurred in laboratories, which not only provide physical proximity to scarce instruments and other scientists but also serve as social organizations for the dissemination of knowledge and training of future scientists [13, 23]. More recently, scientific research teams have engaged in geographically dispersed group work using the CSCW tools included in collaboratories.

The collaboratory is a new organizational form that uses electronic facilities to bring together scientists, instruments, and information to support the conduct of distributed science and engineering work [12]. Designing the systems to support this work requires a careful understanding of current practice and the needs of the research teams involved [33]. One important component of modern scientific work is the collection, analysis and sharing of data. We believe that designing CSCW systems to support the use of scientific data demands an understanding not only of the nature of the data themselves, but also the practices they represent and the functions they serve. We further believe that the near future will bring an increasing demand for effective data-sharing systems, given recent global trends in academic research [4, 5, 37].

Others have taken a similar stance on the use and sharing of documents. Malone's work on how people organize their desks, for example, proved influential in the design of today's desktop computers [31]. It has also been suggested that documents have a "social life," and that they serve multiple roles in many facets of social structure [8, 20]. Additional studies focus on how people manage their paper documents, with an eye toward the design of digital information management systems [39, 44].

As with electronic document systems, creating digital data repositories is a non-trivial problem involving more than just providing remote search and retrieval functionality. We argue, however, that data are different from documents in important and fundamental ways, and warrant separate study. Specifically, scientists regard data as accurate representations of the physical world and as evidence to support claims [27]. As a result, data play several unique and important social roles within research teams. In this paper, we explore the nature of these roles and discuss their specific implications for the design of effective data-sharing systems.

2. DATA SHARING: IMPORTANT, BUT DIFFICULT
Data sharing is important for two reasons. First, data sharing has historically been considered a hallmark of modern scientific practice [32]. Openness in the scientific process allows for the confirmation of research findings, especially through the replication of results [25]. Data sharing also makes it possible for scientists to build on the work of others [30]. It is with this in mind that scientific funding agencies are beginning to require grant recipients to share the data produced in their studies [2, 3].

Second, new "big science" projects involve data that are collected and analyzed by multiple people, institutions, and research sites. Sharing data in these cases becomes more than just the exchange of finalized data sets. Cyberinfrastructure [5], collaboratory, and e-science [37] initiatives in the United States and the European Union are looking at ways to allow scientists to collaborate on the creation and use of very large data sets.

GROUP'03, November 9–12, 2003, Sanibel Island, Florida, USA.

2.1 Difficulties in data sharing

Despite this importance, however, sharing data is not easy. Many researchers have discussed the problems underlying this seemingly simple process [10, 22, 30]. We divide these problems into three categories: 1) willingness to share, 2) locating shared data, and 3) using shared data.

First, there is a strong sense in which the scientist's ability to profit from data collection depends on maintaining exclusive control over the data—economists would say that the data are a source of "monopoly rents" for the scientist. In this case, however, the profit, or rent, accrues largely in the form of scientific reputation and its accompanying benefits, such as publications, grants, and students [43]. The point here is that the competition for reputation (and associated benefits) in science is intense, and there may be a strong reluctance on the part of scientists to share data, as such sharing may amount to a sacrifice of future rents that could be extracted from the data were they not shared [41]. It is also the case that certain data sets funded by private commercial interests may carry usage and confidentiality restrictions that prohibit them from being shared.

Second, researchers must become aware of who has the data they need or where the data are located, which can be a nontrivial problem [45]. After finding appropriate data, they often must negotiate with the owner or develop trusting relationships to gain access [40].

Third, once in possession of a data set, understanding it requires knowledge of the context of its creation [18]. How was each datum collected and analyzed? What format are the data in? If the data are in electronic form, is there a key or metadata available to indicate what the various fields in the database mean? Researchers also need to know something about the quality of the data they are receiving, and if the original purpose of the data set is compatible with the proposed use. Answering these questions in useful ways requires a large amount of effort on the part of the data creator, but the benefit of such effort goes largely to the secondary user. This renders it unlikely that adequate documentation will be produced [17, 34].

Even if documentation is provided, however, it is often the case that much of the knowledge needed to make sense of data sets is tacit. Scientists are not necessarily able to explicate all of the information that is required to understand someone else's work. Collins' [11] discussion of the difficulties in replicating the TEA laser is suggestive here. Collins found that this specific type of laser could not be replicated in different labs simply by following explicit written instructions. Successful replication required extensive contact with someone else who had already built a TEA laser. Knowledge transfer in this instance is not simply a matter of sharing a set of instructions, but is a highly social process of learning practices that are not easily documented.

2.2 Approaching this challenge

Data sharing systems have developed a number of approaches to these issues. For example, standardized reporting formats and metadata protocols can allow the same data to be read across different hardware or software. Password and security systems can give a certain degree of control over who does and does not have access to data sets. Metadata can provide context about who collected data and how they were processed.

While these approaches deal effectively with the explicit technological problems inherent in data sharing, it is not clear that they adequately deal with many of the tacit and social issues outlined above. For example, metadata can provide information about the data, but it too relies on a good deal of insider knowledge to interpret. Indeed, a metadata model can only provide so much contextual information, leading to a potentially recursive situation in which metadata models require "meta-metadata" in order to be effectively understood [6, 40]. Similarly, access control systems implemented for data can put up a wall around a data set, but they do not adequately provide for the subtle social realities of gaining access to an invisible college [35].

Recent calls for open science and data sharing suggest that funding agencies believe that groundbreaking scientific research requires more data sharing among scientists. Even if we provide the technical means to move data from one lab to another, however, there may be social barriers to effectively using this data in practice. To design technologies that truly support the conduct of science, and not just the sharing of a data set, we argue that the designer must understand both the scientific role that data play in producing knowledge, and the social role that data play in the conduct of scientific work.

Before proceeding, we should also note that there are community data systems in current use that are quite effective. Resources such as the Inter-university Consortium for Political and Social Research, Genbank and the Protein Data Bank are predicated on the ability to share data among scientists. We argue that these are special cases, however, in that they represent research areas that are characterized by what has been termed low task uncertainty, and high mutual dependence [15, 43]. In other words, these research areas have a high level of agreement on the types of problems to be studied and the methods to be used. The problems being addressed are also sufficiently large that researchers are dependent on large groups in order to tackle them effectively. In contrast, the research areas that we study here feature a wider array of questions, methods and data formats. This variety complicates the data-sharing problem.

3. METHOD

We begin this work with no assumptions about the purpose of data, but instead draw on our observations of practice in different scientific disciplines. We attempt to develop an understanding of the use of data in practice that can inform the design of data management and sharing systems. We acknowledge both scientific and social understandings of data's purpose and use with the explicit belief that understanding multiple points of view will help us to create better technologies.

In our work developing and designing collaboratories [12], we have actively studied scientists in three disciplines: earthquake engineering, HIV/AIDS research, and space physics. We recognize that these disciplines differ along several important dimensions, but here our goal is to look across the disciplines to explore commonalities and differences in the ways data are used.

Earthquake engineers are interested in the effective design of structures capable of withstanding substantial seismic forces. They test model structures under simulated earthquake conditions in a laboratory or field environment. Data collected include photographs of structural damage, measurements of cracks in concrete, and readings from sensors attached to the specimen. Our involvement with this community has been ongoing since 2000, as part of our work on the U.S. National Science Foundation's (NSF) George E. Brown, Jr. Network for Earthquake Engineering Simulation (NEES). During this time, we have visited 15 earthquake engineering laboratories to observe day-to-day work and experimentation, conducted over 70 half-hour interviews with

faculty, students and technicians, and administered 3 surveys to samples of 400 academic researchers and commercial practitioners.

Clinical and experimental HIV/AIDS researchers are involved in research to understand how the HIV virus functions, how the human immune system responds to the HIV virus, and to test the efficacy of various treatments. Typical sources of data in this field can include laboratory experiments, blood analyses, clinical descriptions of patient symptoms, and epidemiological data about disease incidence in a community. We have been involved with HIV/AIDS researchers for two years. We have visited 9 labs and clinics in 3 countries in Africa and North America, conducted over 25 hours of observation of day-to-day lab work, and 50 one-hour interviews with researchers, students and lab technicians.

Space physicists focus on the study of the earth's ionosphere, looking particularly at the interactions of the solar wind, the earth's magnetic field, and the characteristics of the upper atmosphere [14]. Data are collected from a various ground- and space-based instruments measuring solar radiation, auroral activity, and the like. In this area, members of our research group were involved for over 10 years in the creation and operation of the SPARC/UARC collaboratory. Work included interviews, observations and a trip to the Sondrestrom Research Facility at Kangerlussuaq, Greenland to observe a data gathering campaign.

4. HOW DATA CONTRIBUTE TO SCIENTIFIC FACT

In this section, we focus on the more obvious uses of data in the scientific enterprise. Here we identify three aspects of the way data are used that will impact how data get shared.

4.1 Data as News or Confirmation

Expectations about what purpose data will serve influence how they are collected, arranged, maintained, and shared. In some cases, data serve as confirmation of scientific expectations. In medical research, for example, clinical trials are designed not to generate new theory, but to confirm whether or not an existing theory (for example, about how a drug works) is correct. On the other hand, data can also function as a kind of "news." Investigators are not sure what they will find, or are looking for cases that may push the boundaries of theory.

In HIV/AIDS research, data serve both purposes. In some studies, especially those considered to be "basic research" or "bench science," investigators are often looking for news from the data. An important research question in HIV/AIDS work involves understanding how to predict if or when a person who is infected with the HIV virus will begin exhibiting the symptoms of the disease AIDS. One prominent theory suggests that there is an inverse relationship between the amount of HIV virus in the blood and the strength of the immune system. HIV-positive patients with weak immune systems are more likely to convert to full-blown AIDS.

Generally, there are two research approaches to this question, which are not necessarily incompatible: The first, confirmatory, approach tries to prove the theory by gathering enough evidence to statistically demonstrate that HIV viral load and immune cell counts have a negative correlation. The second approach tries to refine the theory by locating and looking carefully at the small subset of HIV patients who do not fit the prevailing model. These patients, called "non-converters", can have high levels of the HIV virus, but their immune systems remain strong. It is these unexpected cases that carry the most information in this situation, such that the data are treated as "news" about poorly understood aspects of the disease, rather than as statistics to prove a theory. Additionally, when researchers are searching for news rather than confirmation, they are more likely to alter their methods or focus and concentrate on those areas that are most news-rich. Scientists involved in a confirmation study cannot change their methods without jeopardizing the validity of the study.

In earthquake engineering, we find that most medium and large-scale experiments are confirmatory. Because these experiments often require months of preparation and cost tens to hundreds of thousands of dollars, it would be highly impractical to run a series of tests with the goal of learning from anomalous cases. Rather, it is typically the case that researchers work before the experiment to develop a sophisticated numerical model of the physical phenomenon being tested. This model is then used to inform the design of the experimental specimen and apparatus, and the experiment itself is used to validate and refine the theory behind the numerical model.

It is, of course, the case that there are times when data function as news in earthquake engineering experiments. As one faculty member put it, "This is research. If the unexpected never happened, we wouldn't learn anything." In cases where the specimen fails, or collapses, before it is expected to, for example, further study must be done to understand why the unexpected failure occurred. Unlike AIDS research, however, we never found a case where encountering the unexpected and learning from it were the stated goals of the research.

4.2 Data Streams vs. Data Events

One of the differences across the disciplines we are studying is when and how data are collected. In many ways, the timing of data collection influences the rhythms of production and work within the field.

In HIV/AIDS research, data tend to be collected in relatively constant streams over long periods of time. For example, in a clinical study, each patient will have blood drawn and analyzed at regular intervals (perhaps once per month), and could be participating in the study for several years. There are a large number of patients in the study, however, such that each day a relatively regular number of blood samples to be analyzed are produced by the clinic. One or more data points are extracted from each of the blood samples, so that over the life of the study, the day-to-day collection of small amounts of data results in a huge data set.

A potential difficulty for studies that use streams of data is that the context for data collection may change as the study progresses. One HIV clinical trial we studied has gone through three major protocol changes mid-stream in response to both internal and external forces. Access to new and better medicines means that baseline immune responses in study populations have improved. Findings from other studies create an ethical obligation to provide patients in the study with a different combination of drugs. Internally, the project has been forced to deal with staff changes, new laboratory techniques, and the use of different laboratory and clinical sites. Each of these changes potentially requires a new interpretive framework for the data.

Earthquake engineering, on the other hand, tends to collect data from discrete events rather than streams. Several months are typically spent planning and constructing a specimen, instrumenting it with a variety of sensors and preparing it for a simulated

earthquake. The physical simulation itself, referred to as a "shaking event," may last only a few seconds, during which data are collected at a very high rate. A typical laboratory investigation might involve the analysis of data from five or six shaking events performed on a particular specimen, where each event is treated as a discrete earthquake of a particular magnitude and ground motion. These data are generally used solely for the investigation from which they were collected, and are not combined to form a larger data set.

The space physics community has an interesting approach to data in this regard. The data collected in space physics are observational data of natural phenomena produced by the sun. Data could be collected in a constant stream over time, but the field sets artificial boundaries to create discrete events. Rather than always collecting data, a "campaign" of one to two weeks is organized among the researchers when an interesting moment is expected to occur, and multiple instruments are set up to capture appropriate data during the same time period. In some ways, this bounding of events out of the streams may be an artifact of the practical difficulties of data collection. Because the instruments are located in remote areas (like Greenland), scientists would have to schedule trips to visit the instrument and collect data. We may see a change in this pattern as more of the instruments in space physics are connected through electronic networks. Because researchers will no longer be limited by when they can travel to the instrument, we might expect the field to move toward a more stream-oriented data collection approach.

5. HOW DATA CONTRIBUTE TO SCIENTIFIC COMMUNITY

If we accept that science is a social enterprise [16, 28], we must also ask how data are implicated in this social world. For this, we draw on the concepts of "legitimate peripheral participation" and "communities of practice" [29, 42]. The concept of "community" allows us to see not just the static existence of relationships within a group of researchers, but also the processes by which relationships between people and objects at multiple levels of analysis form, change, and cease to exist. Wenger notes that for any community, membership and status are fundamental issues. Focusing on scientific communities of practice allows us to ask how data are implicated in the formation and maintenance of communities of practice.

5.1 Data define boundaries between communities of practice

It seems clear that in many cases, different scientific communities use data in different ways. A fundamental distinction in many fields occurs between experimentalists and theoretical modelers. This distinction has been well developed, for example, in studies of the high energy physics community [38]. Experimentalists directly collect empirical data and use these data to test theoretical hypotheses about physical phenomena. Theoretical modelers, on the other hand, develop sophisticated numerical models of physical phenomena, and compare the output from their models with empirical data for validation and refinement.

Though these groups are distinct, it is clear that there is a symbiotic relationship between them. Modelers need empirical data to validate their models, and experimentalists benefit from the advanced theory that modeling enables. It is also clear that these groups are distinct within the earthquake engineering community, for example. When asked about their research interests, nearly every interview subject indicated early on that they did primarily experimental or theoretical work. Several faculty members we spoke with also indicated that they notice that students tend to be particularly good at one type of work or the other. It is "extremely rare" to find a student who is good at both.

We also noticed different attitudes and approaches toward data in these groups. For theoretical modelers, data serve as both a starting point for model development and an external benchmark for the models. They are often frustrated by the difficulty of obtaining and understanding data sets. For example, one subject noted that experimentalists frequently claim they cannot find data when he requests it, and that even when he does receive data, they are often in a format that is difficult to decipher. Theoretical modelers also tended to view data as more public, seemed more interested in the data-sharing capabilities of emerging collaboratories, and were more upset about failed data sharing attempts in the past. Experimentalists, on the other hand, were generally more possessive about data they had worked hard to generate, and were less willing to share.

Keeping all of this in mind, it is also interesting to note that large investigations in earthquake engineering increasingly involve collaboration between faculty from both groups, and that the experimental investigations are used to further development of an advanced theoretical model. In another example we are familiar with, a group of senior space physicists who do primarily empirical work have cultivated a small team of younger scientists who do modeling work. The senior scientists are largely dependent on the modelers for validation of their theories, and the modelers are dependent on the senior scientists for their funding and data to validate their models. In both of these fields, experimentalists and modelers are increasingly negotiating collaborative relationships around larger projects that satisfy the needs of both groups, rather than working separately and trying to develop a sharing relationship after the fact.

There is a strong sense in which the observed distinction in attitudes makes intuitive sense. Empirical data are used differently by the two groups of researchers. For empirical work, data are the output of an experimental or observational process, which involves substantial effort on the part of the experimenter, and then serve to support hypotheses or provide a basis for further explanation. In theoretical modeling, on the other hand, data are used as input. The modeler begins with, and cannot do his/her work without, the experimentalist's observations of natural phenomena. Thus, it is not surprising that the experimenter wants to extract maximal rents from his/her efforts, while the modeler feels that the raw observations forming the basis of his/her research should be freely available.

Though this distinction in data usage is fundamental to fields like earthquake engineering and space physics, a difference in the way data are used does not always correspond to membership in different communities of practice. In AIDS research, the same scientists are often involved in both basic bench science and clinical trial research. Even though one type of research is using data as news and the other is treating data as confirmation, this does not create a distinction between separate communities of practice.

On the other hand, competition between research projects is intense in HIV/AIDS research. Often two research groups will be trying to answer the same question using the same methods, and they are essentially in a race to be the first to make critical discoveries. The competition in the early years of the epidemic to be recognized as the discoverer of HIV has even been chronicled in the popular media [1]. This is an extreme case, but our informants told us that

the anxiety about getting "scooped" on a discovery is keenly felt. While HIV researchers are often quite open about sharing laboratory techniques between labs, unpublished data are almost never shared. In a discussion about the possibility of multiple projects sharing the cost of developing collaboratory facilities, maintaining the data boundary was a top priority. "They won't be able to see our data, right? Because they are competitors, and we are trying to beat them [to publication]." Similarly, laboratory technicians told us that while they are free to share data within their project, to share data outside of the project requires the consent of the lab director. Data are a primary asset for the scientists, and being able to fully exploit the value of the data demands that the boundaries between communities of practice be maintained.

Other distinctions that we have not fully discussed here can also separate communities of practice. For example, experimental vs. field data, or quantitative vs. qualitative data are distinctions that often define the differences between disciplines and subdisciplines.

5.2 Data as gateway into communities of practice

Access to data can be an important point of entry for a community of practice. In some fields, such as earthquake engineering, the data necessary to engage in scientific activity can be generated in a local laboratory. In other fields, like space physics, the data come from large, remotely located instruments. To become a space physicist requires access to either a data source or to someone else's data set. Instruments, such as the Greenland facility or satellite-based devices, are typically owned by institutions and scientists at these institutions have primary access. Authorship on publications is often exchanged for access to data from such instruments. We have also observed anecdotal evidence to suggest that relationships are negotiated around data access in this community. At a conference, one graduate student we were talking with abruptly left our conversation so that he could go "schmooze [with a more senior scientist] for data."

In earthquake engineering, we spoke at length with several researchers about their attitudes toward data sharing in the NEES collaboratory. One theme that emerged repeatedly was a willingness to share a low-resolution abstraction, such as a graph of experimental data immediately, but significant reluctance to share the full data set. For example, many researchers indicated that they would be happy to allow any interested parties to "tune in" to a live experiment and view graphical displays or summaries of data , but they did not want others to have access to the numerical data files for at least six months. In this way they are able to retain control over their data and potential findings, in addition to influencing the degree to which outsiders can become involved in their research and the field as a whole.

In another example, this one in HIV/AIDS research, data often cannot be understood without a thorough understanding of the clinical and laboratory processes that produced them. For instance, an Elispot test is used to determine whether an individual's immune system is responding to particular portions of the HIV virus. White blood cells and sections of the virus are mixed in a small well with various chemicals and dyes. If the white blood cells recognize the piece of the HIV virus as a threat, they secrete a chemical that reacts with the dye to produce a dark spot on a reaction plate. These dots are counted, and a determination is made as to whether this number of reactions constitutes a positive response to the virus. The number of dots and the amount of background noise (that is, dots that appear but are not due to a white blood cell reaction) are highly contextually dependent. The number of dots that must appear before achieving a "positive" response depends on the individual's and population's general immune health, the procedures used to prepare the reaction plates, the amount, type, and quality of the reagents, and the general level of quality control within the lab where the test is conducted. The important piece of data—the number of dots that appear—cannot be interpreted, and the experiment cannot be replicated without an understanding of all of these (and probably more) factors (for a similar example, see [9]).

This degree of contextual dependence means that giving access to data requires more than merely handing over a data set. To use the data set, the receiver must often visit the laboratory or clinic where the data were generated and must spend time with the researchers to understand what the data mean. In busy research labs, these visits can be both time-consuming and disruptive to the day-to-day work. One lab has formalized this process by requiring any visiting researchers to submit an application that details not only their own research, but how they can give back to the lab (often by providing training or seminars). "We want to know that they have a legitimate reason for being here, and we want to know what we are going to get in return [for letting them in]."

Thus, access to both data and their surrounding tacit knowledge become valuable resources for those seeking scientific legitimacy. Those senior researchers who determine who can and cannot have access to these resources effectively become gatekeepers of the field.

5.3 Data as indication of status in a community of practice

In the scientific communities we are studying, data often indicate status. Having one's own data is often seen as "better" than using publicly available or borrowed data sets. This is highly significant in that it illustrates that some fields have not just a reluctance to share data, but a reluctance to use other people's data as well.

For example, one of the proposed benefits of the NEES collaboratory is the ability to share data sets from large experiments, so that many experimental researchers may analyze them. The data sets from these experiments are frequently much larger than necessary to study the phenomenon specifically being investigated by the researcher, as the cost of adding additional instruments and sensors is low once the specimen has been prepared. Interestingly, many of the researchers we spoke with are interested in making their data public and sharing it, but no experimental researchers indicated a desire to analyze data from other researchers. Some informants suggested that there is a stigma in the community associated with using data collected in somebody else's experiment for one's own analyses. This stigma can affect one's chances at getting papers accepted in journals and conferences, as well as the respect that is accorded by the community.

In space physics, there is also a sense in which possession of data by virtue of institutional access to a data source affords a certain status and enhances one's reputation in the community. One who has ready access to such data, for example, does not have a need for co-authors and has the clear upper hand in negotiations with more junior scientists requesting access to said data.

In HIV/AIDS research, data quality is an important status marker. Obviously, the quality of data is important for the ability to make scientific conclusions, but assessments of data quality often function as commentaries about the researchers themselves. In newer labs,

producing high quality data is a signal that the lab had reached a certain level of maturity. Especially in laboratories located in Africa, we often heard comments like, "Our data is as good as any lab in Europe or the United States." We heard a good deal of boasting when an external review board rated a project's data quality as "excellent." Anecdotal evidence suggests that a discovery of low data quality in a published study resulted not only in disputes about the scientific merit of the work, but also in a loss of status in the HIV/AIDS research community for the associated researchers.

In many ways, these examples make sense in light of what Fuchs [15] and Whitley [43] refer to as the level of task uncertainty in a research area. Task uncertainty refers to the degree to which the researchers in a given area agree on the problems that are to be solved and the methods that should be used to solve them. In areas with low task uncertainty, such as high energy physics, it stands to reason that the stigma attached to using shared data will be lower than in the fields we are studying because the experimental apparatus and means for data collection, and the community faith in these are not likely to differ substantially between researchers. Where there is high task uncertainty, on the other hand, there is more variation between researchers. Because of this, much of the creative effort involved in the conduct of research goes into the design of the experiment itself and conceiving of novel ways to collect data. Simply analyzing somebody else's data bypasses this step, which arguably "counts" for less in the battle for scientific reputation.

5.4 Data enable inbound trajectories

As noted above, communities are dynamic groupings of individuals. Wenger argues that communities of practice are comprised of individuals on various types of trajectories [42]. One type, the inbound trajectory, is characterized by individuals on their way to core participation in the community. Inbound trajectories are made possible when non-members or peripheral members of a community of practice are given opportunities for full participation.

We have observed that possession of data can be a powerful tool for those on inbound trajectories in certain scientific communities of practice. In AIDS research, for example, a graduate student or junior researcher is often given responsibility for a subset of the data on a study. They will have to collect, maintain, and analyze that data on their own. In return, they get valuable mentorship from senior researchers, learning not only how to manage data, but also how to be a member of this community of practice. Even though the junior scientists remain outside the core of the community—they cannot initiate studies on their own, and they do not have ultimate responsibility for a study—this arrangement gives them the opportunity to learn how to interact with funding agencies, ethics review boards, and other scientists. They also gain some of the status markers of community membership, especially authorship on publications that result from their data. This is a kind of "legitimate peripheral participation," in which students are given a real opportunity to act as a part of the community [29]. The work of collecting and managing data provides a peripheral member the opportunity to move closer to the center of the community of practice.

The same does not appear to be true in experimental earthquake engineering. Here, it is largely the graduate students who are at the core of research practice. Rather than having a student assigned to a small "slice" of a larger experiment, each student is typically assigned to his or her own experiment which is worked on with a faculty advisor. The student supervises (and often participates in) the construction of the specimen, the running of the tests, and is solely responsible for the analysis of the data. This responsibility is rarely shared, and we observed no cases where junior students or undergraduate lab assistants were permitted to assist in the analysis. Thus, for earthquake engineers, data are the privilege enjoyed primarily by core (and nearly so, in the case of advanced graduate students) participants in the community who have run their own experiments. Those wishing to establish an inbound trajectory may do so through laboratory work, but not through data analysis.

This distinction between AIDS research and earthquake engineering is critical in understanding willingness to share data. In AIDS research, sharing data with a new student is a typical way to give them some experience and bring them into a community. In earthquake engineering, on the other hand, these opportunities for legitimate peripheral participation occur in the planning and execution of the experiment itself. Thus, willingness to share data in this community is likely to be lower.

6. DISCUSSION

We have seen here that social issues frequently act as barriers to data sharing in ways that technical systems will not soon be able to change. At the same time, however, there are opportunities to use technical systems to support social behavior in ways that can facilitate the sharing of data. We believe that our observations have implications for two issues in the design of CSCW systems to support the sharing of data in science and engineering work in fields such as the ones we studied: 1) promoting sharing behavior, and 2) considering context. We conclude this section with a set of specific recommendations.

6.1 Promoting data sharing

One way to think about data sharing involves economic incentives. Grudin [17] argued that people will be unlikely to use a system if it requires them to do additional work from which they will not benefit. It stands to reason, then, that they will be even less likely to use a system which requires extra work and has the risk of a negative payoff.

We noted above that scientists with their own data have the opportunity to extract monopolistic "rents," or revenues, from these data in the form of reputation, publications, and such. Sharing these data often amounts to risking the loss of these benefits. At the same time, however, because data are regarded as representations of the external world, it can be said that many data sets incorporate a larger set of potential rents than the creator is capable of or interested in extracting. After all, scientists can only do things with their data that they are capable of, have thought of and are interested in doing.

From this, we can divide the theoretical full set of potential revenues that the creator of a given data set can extract into four categories. Consider the hypothetical example of Jane the earthquake engineer. Jane has sensor data from recent laboratory tests of two concrete columns, each built with a slightly different concrete composition. Uses of these data can be categorized as follows:

(1) "A scientist's data set is her castle" includes revenues from activities the creator is capable of and plans to do. In Jane's case, this might be a basic statistical comparison of the performance of the two columns under simulated earthquake conditions.

(2) "With a little help from my friends" includes revenues from activities the creator has thought of and wants to do, but seeks collaborators with the necessary skills to do them ([19] lists

this as a central reason for much collaboration in science). For example, Jane may wish to run more sophisticated statistical analyses that are beyond her ability. In order to do this, she seeks a collaborator with the necessary skills.

(3) "One scientist's junk is another one's treasure" includes revenues from activities not relevant to questions of interest to the creator. In this case, Jane is interested in the performance of the two types of concrete, and has very little interest in the steel reinforcement bars (rebar). She has sensor data from the rebar, however, that she collected in case of unexpected behavior. She would be happy to give these data away to anybody interested, because they do not help answer her core questions.

(4) "D'oh!" includes revenues from activities the creator has not thought of, but that are relevant to questions she finds interesting. Jane is a good earthquake engineer, but she does not know everything. If it turns out that there are some basic statistical analyses of the columns' performance that she did not think of but that reveal interesting discoveries, it could be quite embarrassing to Jane to have these pointed out by an outsider analyzing her data.

If we assume that people are honest and give credit where it is due, we can suppose that the reputational payoffs of types 1 - 3 are all positive, and that they are listed in descending order of value. Category 4 activities carried out by someone other than the creator are potentially embarrassing, however, and could have a negative impact on the creator's reputation.

Now think about how these categories of revenues might impact a researcher's propensity to share data. The risk of losing category 1 and 2 revenues is arguably too great to warrant sharing one's raw data publicly with the hope of accruing category 3 revenues (in the form of possible co-authorships and acknowledgements). Additionally, there is a risk of embarrassment from category 4 activities that might preclude immediate sharing of data with colleagues interested in similar questions. Instead, as we (and others) have observed, scientists seek to control who has access to their data and share it only with those whom they trust after they have finished preliminarily analyzing it.

Notice that one common feature of the "barrier" data roles we illustrated is that virtually all involve social relationships. Researchers frequently wish to limit access to their data and communities to those with whom they already have relationships. In other words, there is a sense in which the sharing of data follows the paths established by existing social networks

Thus, one possible way to encourage data sharing behavior may be to provide facilities for communication around shared data abstractions. With data abstractions that give a sense of what is captured in the data without giving away too much detail, such as the graphically represented data from live earthquake engineering experiments mentioned above, the risk of losing category 1 revenues can be reduced to the point where the possibility of accruing category 2 and 3 revenues provides sufficient incentive to share these abstractions.

In this scenario, note that category 2 revenues become a possible benefit, rather than a risk, in that the creator is arguably more likely to find a collaborator for category 2 activities by sharing an abstraction than by sharing nothing at all. In this way, the abstractions become a signaling mechanism within the data system in that they provide a limited set of information about what researchers in a particular lab are working on. These signals, in turn, serve both to reinforce the community-of-practice boundary role of data that we observed by limiting access to the full data sets, while at the same time allowing researchers to tentatively reach out across these boundaries and seek out new collaborators.

This strategy may also help reduce the risk of hidden category 4 activities that could result in reputation loss. Sharing preliminary data abstractions gives the creator a chance to receive comments and questions that may reveal additional avenues of inquiry before they become embarrassing. The creator can then turn a potential "D'oh" into a category 1 or 2 revenue.

One implication of signaling one's activity via abstraction to a larger community is the potential need for filtering of incoming communications. Indeed, not all incoming requests for data will be from competent or serious potential collaborators. Thus, there will likely be value for automated filtering mechanisms to assess the seriousness and utility of such incoming messages.

Resnick [36] suggests that using technical systems to build value in social networks results in "sociotechnical capital." We are suggesting that by sharing data abstractions and intentionally leveraging the power of communications tools, scientists can build relationships with potential collaborators. At first, these relationships will facilitate data sharing. After that, they will facilitate the sharing of important aspects of the context of data collection and interpretation.

We are not suggesting that merely adding communication tools will remove all barriers to data sharing. Indeed, there are already a large number of communication tools available to scientists who wish to share data, such as email and the telephone. Instead, we believe that these tools combined with data abstractions can facilitate the development of social relationships that surround the important roles that data play in research communities. Specifically, sharing data abstractions and providing means for communication around these abstractions could allow more people access to the "gist" of the data, and perhaps inspire a student with a novel idea to contact the data's creator. They could discuss this idea, and if it is truly novel, the creator could elect to share the full data set. This can further serve to reduce barriers to entering the core of a community [21, 24].

6.2 Considering context

We believe that data standards have been difficult to establish for two reasons: 1) they do not consider the roles that data play in research communities, and 2) metadata models are not as simple as they seem.

In the first place, those developing community standards for data repositories and the like must carefully consider the roles that data and data standards play in that community. In their analysis of the International Classification of Disease (ICD), Bowker and Star [7] note that the numbered labels comprising the ICD are part of a larger "genre system" that includes codification practices and nomenclature lists. This genre system, in turn, is embedded in a large and complex social context. Disagreement between doctors on what constitutes various diseases (for example, when does a patient move from being merely HIV-positive to "having AIDS") is common. In this way, the system is not foolproof and is not viewed as a complete or static solution. It nonetheless serves an important function by providing a stable classificatory reference point within the community.

Similarly, we argue that those developing community data standards should not seek to create a complete solution that pleases every

researcher in the field. As research fields evolve, so too will the data standards. We have suggested that data are not simply carriers of meaning, and that converting raw data into scientific or social meaning is an active, context-dependent process. Data cannot be easily handed off from one scientist to another. Understanding what data mean is more than just a simple process of re-running some analyses. While metadata can help make sense of data sets, metadata alone will not resolve these complexities. Indeed, the metadata model for one project we are involved with includes a possible 297 fields for each individual data point! Even with this level of detail, however, several of our interview informants suggest that this model is not sufficient to fully understand what took place in an experiment.

Rather, we argue that it is more important to understand the roles that data play in communities and to support sharing the crucial contextual bits of information that data users will need. For example, an archival system for a "stream"-based research community will likely look quite different from one designed for an "event"-based community in that the context of data collection is vastly different. While understanding the data for an EE experiment, requires a detailed description of the experimental apparatus at the moment of testing, , understanding HIV/AIDS study data requires knowledge of the study population and how it changed over time. Some of this can be captured through documentation, but conveying certain tacit components might benefit from the communication tools mentioned above.

There is also a crucial distinction in how data and contextual information might be stored for communities that use data to look for news, as compared with communities that use primarily confirmatory studies. This distinction is particularly important in considering the way in which anomalous data are treated. For confirmatory research, anomalies are generally considered to be "noise" and are often deleted or ignored as part of the standard data "cleaning" process that precedes analysis. In "news" research, on the other hand, these anomalies are at the center of the investigation and have the power to shift the nature of the research. In this way, communication of the context of data processing also emerges as central to the sharing process.

6.3 Design recommendations and future work

We have identified three specific CSCW design recommendations that stem from this work:

(1) Support social interaction around data abstractions and the data themselves – We have shown that data sharing in the fields we have studied requires existing relationships in which sharing takes place. Understanding others' data also requires the sharing of tacit knowledge about the creation and interpretation of the data. The significant recommendation here is that of shared abstractions. By enabling shared data abstractions, we believe the propensity to build relationships to foster sharing will be increased. Once these relationships are established, CSCW tools for data sharing must support interaction between those who created the data and those who wish to use it.

(2) Do not rely on metadata alone – Metadata models are useful for understanding how data were collected, but they are an inherently incomplete abstraction of what actually took place in an experiment. We believe that it is also necessary to support the sharing of supplementary materials that enhance the value of the data. Such materials might include indicators of data quality, design documents for an experiment, and clear indicators not only of what the data represent, but who collected them. Data have many meanings, and reputation matters a great deal. By supporting the sharing of information that enhances the value of data to both creator and user, the actual uses of data are better supported and sharing will be more likely to occur.

(3) Support social and scientific roles of data – Existing data sharing systems focus heavily on support for scientific uses for data. We have illustrated that data serve a variety of other roles in communities of practice, and suggest careful observation as part of an iterative design process, so that all of these roles may be effectively supported.

We have further identified two areas for future CSCW research in the area of supporting data sharing for communities of practice:

(1) We need a better understanding of data abstractions. Specifically, we need to know how much information about data can be released without risk of losing what we have referred to as category 1 and 2 rents in various fields. Where earthquake engineers have suggested that graphical representations may be effective in their field, it is likely the case that this will not be the case in all fields or even within fields. Galison [16], for example, distinguishes between the "image" and "logic" traditions of high energy physics which rely on visual depictions of data in qualitatively different ways. Where the logic tradition views images as mere abstractions of "real" numeric data, the image tradition treats the same images as representations of the real world from which conclusions can be drawn directly. We believe that understanding how to represent data in ways that are useful to potential users without being too risky for creators is a major challenge of efforts to support data sharing.

(2) We have argued repeatedly that metadata is not an effective long-term solution for the types of fields we studied in that it fails to capture the context and tacit knowledge of what happened. Metadata protocols further require tremendous amounts of information to be stored for each data point. We believe that metadata protocols have an important role to play in the sense that Bowker and Star [7] describe with regard to the ICD, but that finding more elegant ways to share contextual information is a much harder problem.

7. CONCLUSION

This paper has argued that data serve multiple roles in science and engineering work, and that these roles must be carefully considered in the design of CSCW systems to support research teams. It was shown that data sharing, particularly in fields with high task uncertainty, is a nontrivial problem because of the difficulty of communicating contextual information in the absence of interpersonal interaction. Gaining access to this contextual information, which is often tacit, requires an understanding of: 1) the nature of the data itself, 2) the scientific purpose of its collection, and 3) its social function in the community that created it.

As we write this, we recognize that there are also new forms of data collection and aggregation that are only possible because of advances in computer networks. As the conduct of science becomes larger and projects involve more people, data sets also get larger, responsibility for them is diffused, and the social norms around how data is produced and consumed will change. The ATLAS high energy physics experiment, for example, consists of nearly two thousand scientists who are currently making plans to analyze

hundreds of terabytes of data, which will start becoming available in 2007. Technologists can and should take an active role in the discussions surrounding these projects, but we must also recognize that understanding the roles played by data in these communities is crucial for facilitating effective sharing and collaboration.

8. ACKNOWLEDGMENTS

This work has benefited substantially from comments on earlier drafts by Tom Finholt, Judy Olson, John Walsh and anonymous reviewers. We also wish to thank the members of the space physics, earthquake engineering, and HIV/AIDS research communities who gave us their valuable time. This research was supported in part by the John Evans Foundation, the Waterford Project, and National Science Foundation grants IIS 0085951 and CMS 0117853.

9. REFERENCES

[1] R. Spottiswoode, Dir. *And The Band Played On*, Home Box Office, 1993.

[2] Data Archiving Policy. http://www.nsf.gov/sbe/ses/common/archive.htm.

[3] NIH Data Sharing Information. http://grants1.nih.gov/grants/policy/data_sharing/.

[4] Data and Collaboratories in the Biomedical Community: Report of a panel of experts meeting held September 16-18, 2002 in Ballston, VA. http://nbcr.sdsc.edu/Collaboratories/CollaboratoryFinal2.doc

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

[6] Bowker, G. C. Biodiversity datadiversity. *Social Studies of Science 30, 5* (2000), 643-683.

[7] Bowker, G. C. and Star, S. L. *Sorting Things Out: Classification and its Consequences*. MIT Press, Cambridge, MA, 1999.

[8] Brown, J. S. and Duguid, P. The social life of documents. *First Monday 1, 1* (May 1996).

[9] Cambrosio, A. and Keating, P. "Going monoclonal": Art, science, and magic in the day-to-day use of hybridoma technology. *Social Problems 35, 3* (1988), 244-260.

[10] Campbell, E. G., Clarridge, B. R., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N. A. and Blumenthal, D. Data withholding in academic genetics: Evidence from a national survey. *Journal of the American Medical Association 287, 4* 2002), 473-480.

[11] Collins, H. *Changing Order*. Sage Pubs., London, 1985.

[12] Finholt, T. A. Collaboratories. in B. Cronin (eds.) *Annual Review of Information Science and Technology*. Information Today, Medford, NJ. 2002

[13] Finholt, T. A. and Olson, G. M. From laboratories to collaboratories: a new organizational form for scientific collaboration. *Psychological Science 8, 1* (1997), 28-36.

[14] Freeman, J. W. *Storms in Space*. Cambridge University Press, Cambridge, UK, 2001.

[15] Fuchs, S. The Professional Quest for Truth: A Social Theory of Science and Knowledge. State University of New York Press, Albany, NY, 1992.

[16] Galison, P. *Image and Logic: A material culture of microphysics*. University of Chicago Press, Chicago, 1997.

[17] Grudin, J. Why groupware applications fail: problems in design and evaluation. *Office: Technology and People 4, 3* (1989), 245-264.

[18] Haas, J. K., Samuels, H. W. and Simmons, B. T. *Appraising the records of modern science and technology: A guide*. MIT Press, Cambridge, MA, 1985.

[19] Hagstrom, W. *The Scientific Community*. Basic Books, New York, 1965.

[20] Hertzum, M. Six roles of documents in professionals' work. in S. Bodker, M. Kyng and K. Schmidt (eds.) *Proceedings of the Sixth European Conference on Computer-Supported Cooperative Work*. Kluwer Academic Publishers. 1999.

[21] Hesse, B. W., Sproull, L., Kiesler, S. and Walsh, J. P. Returns to science: Computer networks in oceanography. *Communications of the ACM 36, 8* (1993), 90-101.

[22] Hilgartner, S. and Brandt-Rauf, S. I. Data access, ownership, and control. *Knowledge: Creation, Diffusion, Utilization 15, 4* (1994), 355-372.

[23] Knorr Cetina, K. *Epistemic cultures: how the sciences make knowledge*. Harvard University Press, Cambridge, MA, 1999.

[24] "Synagonism" versus "antagonism": A universal model. http://www.scienceofcollaboratories.org/WorkshopStuff/Nov2002/1-konidaris.ppt.

[25] Krathwohl, D. R. Methods of educational and social science research: An integrated approach. Longman, New York, 1998.

[26] Kraut, R. E., Egido, C. and Galegher, J. Patterns of contact and communication in scientific research collaborations. in J. Galegher, R. E. Kraut and C. Egido (eds.) *Intellectual Teamwork: Social Foundations of Cooperative Work.* Lawrence Erlbaum Associates, Hillsdale, NJ. 1990

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

[28] Latour, B. and Woolgar, S. *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press, Princeton, NJ, 1986.

[29] Lave, J. and Wenger, E. *Situated learning: Legitimate peripheral participation*. Cambridge Press, New York, 1991.

[30] Louis, K. S., Jones, L. M. and Campbell, E. G. Sharing in science. *American Scientist 90, 4* (2002), 304-307.

[31] Malone, T. W. How do people organize their desks? Implications for the design of office information systems. *ACM Transactions on Office Information Systems 1, 1* (January 1983), 99-112.

[32] Merton, R. K. *Social theory and social structure*. Free Press, New York, 1968.

[33] Olson, G. M. and Olson, J. S. User-centered design of collaboration technology. *Journal of Organizational Computing 1,* (1991), 61-83.

[34] Orlikowski, W. J. Learning from Notes: organizational issues in groupware implementation in *Proceedings of Computer Supported Cooperative Work (CSCW)* (Toronto, Canada, November 1-4, 1992) ACM, 362-369.

[35] Price, D. J. d. S. *Little science, big science...and beyond*. Columbia University Press, New York, 1986.

[36] Resnick, P. Beyond Bowling Together: Sociotechnical capital. in J. M. Carroll (eds.) *HCI in the New Millenium*. Addison-Wesley, 2002

[37] e-Science Core Program: Welcome. http://www.research-councils.ac.uk/escience/.

[38] Traweek, S. *Beamtimes and lifetimes: The World of High Energy Physicists*. Harvard University Press, Cambridge, MA, 1988.

[39] Trigg, R. H., Blomberg, J. and Suchman, L. Moving document collections online: The evolution of a shared repository. in S. Bodker, M. Kyng and K. Schmidt (eds.) *Proceedings of the Sixth European Conference on Computer-Supported Cooperative Work*. Kluwer Academic Publishers, Dordrecht. 1999

[40] Van House, N. A., Butler, M. H. and Schiff, L. R. Cooperative knowledge work and practices of trust: Sharing environmental planning data sets. in (eds.) *Proceedings of CSCW 1998*. ACM Press, New York.

[41] Walsh, J. P. and Hong, W. Secrecy is increasing in step with competition. *Nature 422,* (April 24, 2003), 801-802.

[42] Wenger, E. *Communities of practice: Learning, meaning, and identity*. Cambridge University Press, New York, 1998.

[43] Whitley, R. The Intellectual and Social Organization of the Sciences. Oxford University Press, Oxford, 2000.

[44] Whittaker, S. and Hirschberg, J. The character, value, and management of personal paper archives. *ACM Transactions on Computer-Human Interaction 8, 2* (2001), 150-170.

[45] Zimmerman, A. *Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists*. Unpublished Dissertation, Information and Library Studies, University of Michigan, Ann Arbor, 2003.