Beyond Trust and Reliability: Reusing Data in Collaborative Cancer Epidemiology Research

Betsy Rolland[1,2] and Charlotte P. Lee[2]
[1]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M4-B402, Seattle, WA 98109, USA
[2]Human Centered Design & Engineering, University of Washington, 423 Sieg Hall, Seattle, WA 98195, USA
brolland@fhcrc.org, cplee@uw.edu

ABSTRACT
While previous CSCW research on data sharing and reuse has focused on how researchers assess the trust and reliability of the data of others, we know little about scientists' data use practices after that decision has been taken. This qualitative study of post-doctoral researchers' use of preexisting datasets investigates the practices of cancer-epidemiology post-docs working to understand their "Small Data" datasets. We report the ongoing and iterative nature of information seeking inherent in using unfamiliar data and the time-consuming and highly-collaborative process post-docs used to understand aspects of the dataset important to their scientific questions. Understanding data use practices can help inform the design of both Small Data projects and large cyberinfrastructure projects where multi-source data are collected and combined.

Author Keywords
Data sharing; data reuse; collaboration; data; information seeking; qualitative research.

ACM Classification Keywords
H.5.3 [Information Interfaces and Presentation]: Computer-supported cooperative work

General Terms
Human Factors; Design.

INTRODUCTION
Cyberinfrastructure and eScience studies have been of increasing interest to the field of CSCW as an exemplar of the challenges inherent in collaborative work [1,2,11,12]. Such studies have identified data sharing as a key problem within collaborative research. Recent work has discussed the importance of support for data sharing in collaborative science, focusing primarily on the issues of trust and reliability that must be addressed before researchers can comfortably use data collected by others [4,5,7,8,20]. Other work has focused on why researchers share [6] and the effects of sharing on future citations [17,18].

However, there has been little study of how researchers actually use data collected by others once trust and reliability have been established. Without such knowledge, it is challenging to develop systems and practices that support better curation and reuse of existing data. Until we understand both how to prepare data for reuse and to reuse data more efficiently, the challenges of data sharing will continue to slow down collaborative research and data repositories will remain underutilized [16]. A first step toward the goal of understanding the full data reuse lifecycle is to investigate how researchers determine how to use variables from an existing dataset appropriately for their own analyses. In this paper, we report on research that investigated the data reuse practices of scientists, specifically cancer-epidemiology post-doctoral researchers (PhD-level trainees brought in for specific projects) at the Fred Hutchinson Cancer Research Center (FHCRC) in Seattle, Washington, USA, as they worked to understand the datasets recommended for their use by their mentors. In all cases, the mentors had first-hand knowledge of the quality and reliability of the data, either through participation in the original data collection or through their own previous use of the data. This prior vetting by the mentors permits us to move beyond issues of trust and reliability as factors affecting the decision to use the data, allowing us to concentrate on the actual reuse practices of the post-docs.

BACKGROUND
From *The New York Times* to *Nature*, Big Data is a hot topic, while Small Data rarely gets mentioned. In many biomedical research fields, such as cancer epidemiology, small datasets collected for individual studies, when combined or used for new analyses, represent a potential for new discoveries similar to that attributed to Big Data [3,8,9,16]. However, reuse of these small data sets is fraught with challenges. Small datasets can be difficult to find, as they are rarely deposited in repositories, but, rather, live on investigators' local hard drives or lab servers. Documentation is often informal, spotty or nonexistent

CSCW '13, February 23–27, 2013, San Antonio, Texas, USA.

except for minimal details documented in the methods section of published papers. Original study staff may have moved on or forgotten how the study was conducted or how variables were interpreted, creating enormous potential for misinterpretation and misuse of the data. Data curation, the active preservation of data with its accompanying documentation, is rarely a formal process in these small studies.

Such small datasets stand in marked contrast to large, publicly available datasets such as GenBank or government databases with deposit guidelines or standardized data requirements, where documentation is available and vetted (formally or informally) by a community of users, and the infrastructure is maintained by a funded organization of some sort. It is also different than, though related to, data sharing and reuse within cyberinfrastructures (CIs) or large databases constructed in consortia. Such groups spend significant time designing and developing databases to accommodate the different data structures that will be included and generally employ data managers and/or programmers to assist with this process. In order to bring together data from multiple studies into one standardized database, conversations must take place between original study staff and the system developers to ensure everyone involved is clear on the meaning of the variable, the context in which it was collected and assumptions made when coding the variable into its current form. It is these types of conversations that we detail in this paper, albeit in the context of Small Data.

Calls for sharing and reusing data are increasingly loud and urgent, as exemplified by a recent editorial in *The Lancet* by the leaders of prominent funding agencies in the US and UK [19]. The editorial asserts that not sharing research data is harming public health and proposes several steps that the field of biomedical research can take to improve data sharing, encouraging foundations to require data sharing and deposit by all grantees. The US National Science Foundation (NSF) and National Institutes of Health (NIH) have also recently tightened their requirements for data sharing as a condition of funding. Yet calls for deposit and sharing of data rarely address how the eventual recipients of those datasets actually use the data [6]. In fact, we know very little about the practices of researchers who use data collected by others.

Data Reuse: A Definition

The terms "data reuse" and "data sharing" are often used interchangeably. Borgman (2011) notes that calls for "data sharing" are rarely explicit about what they mean by this term, but suggests this possible definition: "… the release of research data for use by others. Release may take many forms, from private exchange upon request to deposit in a public data collection. Posting datasets on a public website or providing them to a journal as supplementary materials also qualifies as sharing" [6]. Data reuse, then, is the work done by the recipient of those shared data. It involves identification of a dataset of interest, receipt of the dataset and appropriate use of the data for analysis.

Advantages and Disadvantages of Data Sharing

There is general agreement that data sharing and reuse can have tremendous benefits for society, if done well. These benefits include the ability to ask larger questions than any individual study might be able to ask on its own, due to a larger and more complex data set; the ability to take advantage of existing datasets rather than spending time collecting new data; and potentially large cost savings because there is no need to assemble new projects. Benefits to individuals include more publications and potentially higher-impact publications, at that, due to the bigger questions being asked.

Such data sharing and reuse are increasingly common in scientific research, in large part due to the development of technological support for doing so. Increased bandwidth and ubiquity of communication technologies allow researchers to share virtually anything with anyone. Furthermore, scientists are culturally accustomed to sharing within their trusted social networks. This stands in contrast to the for-profit world, where data are considered proprietary and rarely sent outside of the walls of the corporation without strict legal agreements.

However, the field of CSCW has documented significant challenges inherent in data sharing. These include difficulties in the interpretation of data and appropriately conveying contextual details of how the data were collected [5]; issues of trust in the quality of data [8,20]; and a reluctance to give away intellectual property [5]. The process of assessing data for reuse can be tricky and time-consuming for both the sharer and the user. However, the cost of an inaccurate assessment of the quality or an incorrect understanding of the data is much higher in terms of both money and reputation. This applies to both the original study collectors sharing their data and the investigator seeking to use it; a retraction of results due to inappropriately used data or the use of bad data is an embarrassment for everyone involved.

Cancer Epidemiology

Epidemiology is the study of disease risk at the population level. Some examples of populations studied by epidemiologists may include post-menopausal women, prostate-cancer survivors, or children. Studies focus on exposures such as tobacco use, family medical history, diet, or exposure to workplace toxins. Cancer-epidemiology datasets possess a number of characteristics that make their use particularly interesting to study. First, epidemiologic data are generally collected by questionnaire. As such, questions must be straightforward enough for a non-scientist to administer and a participant to answer. Although epidemiology questions are often not standardized (although this, too, is an area of change (see, for example, http://www.p3g.org/)), there are some generally accepted

practices that make it easier for researchers to understand one another's data sets. For example, anyone collecting lifestyle data will want to collect smoking data with both frequency and duration in order to produce the derived variable "pack-years," a measure of the amount of cigarettes smoked over time. Additionally, epidemiologists are asking similar questions over different populations. Studying the association between smoking and breast cancer in post-menopausal women is not that different, at a high level, from studying the association of smoking and prostate cancer in Hispanic men. All of these characteristics make for datasets that are easier, if not easy, to combine.

Most importantly, cancer epidemiologists have a history of sharing data within their trusted social networks. This is partially attributable to the similarity of data collected but also speaks to a core belief within the cancer-epidemiology community that a dataset can be explored in multiple ways for a variety of associations. One of the defining characteristics of epidemiologic data is that they are not reproducible, given that they are collected in a specific time and place on a specific group of individuals. Even if another similar group is assembled, the data from the studies can never be identical. This makes epidemiologic data extremely valuable and useful for sharing. The sharing researcher gets fuller use of the data while the user gains access to a gold mine of data without the time and money required to produce them. In times of scarce funding, data sharing and reuse allows science to continue to move forward with fewer funds. By pooling shared data, epidemiology researchers can attack new questions without spending a lot of money.

SITE DESCRIPTION AND METHODS

This research took place at the Fred Hutchinson Cancer Research Center, a National Cancer Institute (NCI)-designated cancer center in Seattle, Washington, USA. This class of research institute is categorized by a high level of scientific excellence and patient-centered research. FHCRC has approximately 3,000 employees and is organized into five divisions. One of these divisions is the Public Health Sciences (PHS) Division, home to most of the organization's cancer epidemiologists. In 2011, FHCRC ranked 19th in total NIH funding, receiving in excess of $270 million, ahead of much larger institutions [15], and 3rd in funding from the NCI. One of the authors of this paper is an employee of FHCRC and works primarily on cancer epidemiology projects, including previous work on one of those described in these interviews. However, the author had little or no prior interactions with any of the participants in this study, meeting most of them for the first time when conducting the interviews.

The National Postdoctoral Association defines a postdoc as "an individual holding a doctoral degree who is engaged in a temporary period of mentored research and/or scholarly training for the purpose of acquiring the professional skills needed to pursue a career path of his or her choosing" [14]. Post-docs at FHCRC are generally brought in to do new analyses on data that have already been collected, either as an extension of the original aims or to undertake a completely different project. They work with a specific mentor, most often the Principal Investigator (PI), the researcher in charge of the project, who originally collected the data under analysis. In some cases, post-doc positions were written into grants and appropriate hires made; in other cases, they were funded under training grants. All post-docs at FHCRC are subject to ethics education requirements and their research projects must receive appropriate Institutional Review Board (IRB) approvals.

We chose to interview post-docs for several reasons. First and foremost, post-docs represent a user group just one step removed from the original data collectors. They must use data someone else collected, with all of the difficulties that entails, yet they still have direct access to those who collected the data, which means access to answers to their questions. Because they were given access to the dataset either by their mentors or via their mentors' professional relationships, they are able to bypass the thorny issues of trust and reliability so well documented by previous CSCW research. Thus, we are able to concentrate on the actual reuse practices of the post-docs. Finally, as junior researchers, post-docs are expected to be learning and developing professional practices while also doing a great deal of work independently; as such, their situation is conducive to reflection.

Over the course of six weeks, we conducted semi-structured interviews with a diverse group of post-doctoral researchers in cancer epidemiology at FHCRC, including four men and seven women. Participating post-docs hailed from several different institutions and fields, including medicine and public health (MD/MPHs), behavioral epidemiology, and molecular epidemiology. They worked with different mentors and had been working in their post-doctoral fellowships anywhere from 3 months to 2 years. Each post-doc was interviewed once, in a conference room at FHCRC, and interviews lasted between 30 minutes and one hour. Participants were asked about all the projects on which they had used preexisting datasets as part of their post-doc, then asked to pick one upon which to focus for the remainder of the interview. The interviews were recorded and transcribed, and all participants were given pseudonyms. While not a full grounded theory study, we did utilize a grounded theory approach to analyze our transcripts. Data were first coded for specific interview questions, then reanalyzed for emerging themes.

FINDINGS

In order to better understand data reuse, we investigated how post-docs determine appropriate use of preexisting data for their own analyses. We asked participants about the process they went through when starting to work with a new dataset and paid particular attention to the information sources they used, as well as their personal strategies for

engaging with the data. In this section, we detail the data analysis process followed by our participants and describe a typology of information needs experienced by post-docs as they went through this process. We then discuss in detail one of these information needs; namely, how the dataset's variables were constructed. Understanding the construction history not only gave post-docs grounding in the conduct of the study but gave them critical information required for their own analyses.

Data Analysis Process

In general, the process post-docs undertook was remarkably similar across participants, despite the fact that they came from a variety of disciplines and universities, worked with different mentors and were working on quite different projects. They described a project proposal phase, during which they worked with their mentors to locate an appropriate dataset and mapped out the project. These project proposals required approval from either their mentor or a full study committee intimately familiar with the data before they could proceed, ensuring a trustworthy, reliable and appropriate dataset for the post-doc's work. Only one of our participants worked directly on the main study database; the rest all received copies of the main database, generally a subset containing only the variables they had requested. The data had been cleaned to varying extents; meaning, the data manager had removed obviously incorrect records or corrected mistaken entries in the database. All post-docs noted they were not allowed to share their dataset with anyone else.

Some post-docs received many derived (i.e., calculated) variables, while others derived their own. For example, the post-doc might receive a variable for "total Vitamin D intake" in addition to data on Vitamin D intake for all the individual food sources asked about on the questionnaire. Deriving variables requires that decisions be made about precisely how to calculate them. Some such calculations are standard, such as Body Mass Index, which is calculated as the weight in kilograms divided by the square of the height in meters. Others may be calculated in a variety of ways according to the scientific thinking of the PI or data manager. In our Vitamin D example, the Vitamin D content of each individual food the participant reported consuming (e.g., carrots, milk, broccoli) will be calculated based on average Vitamin D content of that food according to a government database. Another example might be a derived variable "socio-economic status" which takes answers to questions about education level and family income to create categories of high, medium and low economic status. The study team would make decisions about how much weight to give to each input variable (education and income) and where the upper and lower limits of the categories might be. This scenario could be further complicated by regional differences if done within a multi-site study. High economic status in Manhattan and Ames, Iowa, require very different cutoff points. As we will see later, understanding such decisions is crucial for appropriate use of the data.

Upon receipt of the dataset, post-docs began to get a feel for the data by talking with the study's data manager, running summary statistics (e.g., mean, median, missing, distribution) and checking those against previously published manuscripts. As they performed their analyses, post-docs presented interim results to their mentor and study groups, often getting feedback about their work and their use of the data.

Finally, they wrote up their results into a manuscript for submission to a relevant journal, relying heavily on the comments in their own analysis code to remember the details of the decisions they made along the way. This manuscript was distributed to coauthors for feedback. Throughout this collaborative process, post-docs encountered numerous situations where they needed more information in order to use the data appropriately.

Typology of Information Needs

Post-docs needed answers to a series of questions as they progressed through their project, beginning from the proposal stage through submission of the manuscript for publication. These questions are shown in Table 1. Just as the process of science is not linear, neither is the process of scientific analysis. While the questions generally came in this order temporally, they could arise at any time, and post-docs often looped iteratively and unpredictably through the questions. A colleague's suggestion to utilize an additional variable from the dataset in the analysis could drive the post-doc to return to seeking information about previous analyses which used that variable. It is important to note that questions arose primarily about those aspects of the study that were of greatest importance to the post-doc's current analyses. They were not seeking information for the sake of a vague understanding of "context" of the study; rather, they sought specific details about the study that directly or indirectly affected the work they were doing. The most difficult, time-consuming and interactive questions centered around how and why the data had been constructed the way they appeared in the dataset given to the post-doc. For purposes of this paper, only question 6 about the construction of data will be discussed at length. Other questions will be addressed in more detail in future work.

Question Type	Examples
1. What datasets are available to use in my research?	What datasets does my advisor have access to around my research area?
2. Will this dataset help me answer my research question?	Does this dataset have the population, study design and outcome data I need?

3. What else has been done with this dataset?	What manuscripts have been previously published from this study?
4. Where do I find the information I need to understand this study?	Where can I find copies of the questionnaire and codebook? Which copy of the dataset should I use?
5. How was this dataset constructed?	On what basis did the original study include or exclude participants? Why was this question included on the questionnaire?
6. How were these data constructed?	**How did the original study staff code my variables of interest? Why are so many participants missing data on the variable "alcohol use"?**
7. What do my variables of interest mean?	How did the original study define the variables "heavy smoker" and "frequent aspirin user"?
8. Am I using the data and the dataset correctly?	Does my interpretation of the meaning of this variable match that of previous data users? Did I select the correct data file?
9. What have I done with this dataset?	How did I define my categorical variables in my analysis and how should I represent that in my manuscript?

Table 1: Types of questions asked by post-docs during their project.

Information Seeking Strategies

To obtain the information they needed for their analyses, post-docs in our study employed a number of strategies, taking full advantage of the resources available to them. They started their projects with conversations with their mentors, using the mentor's vast knowledge of existing datasets to identify interesting and available datasets, then narrowing their research interest to a plausible research question for this dataset. Again, questions of trust and reliability were never raised as post-docs assumed their mentors would only point them to good data. Once this data selection had been completed, post-docs began using available written sources of information such as project websites, codebooks and study questionnaires to bolster the information gleaned from previous discussions. Inevitably, questions arose that were not answerable via available documentation, leading post-docs to have further conversations with their mentors and study staff to find appropriate answers. These conversations could lead either to a pointer to new documentation or to new knowledge that didn't exist in any written form, only in someone's memory.

Conversations were occasionally a straightforward recounting of a fact such as what year the study stopped enrolling participants; more frequently, they consisted of unraveling a trail of decisions and interpretations to answer a question such as: Why are so many participants missing tumor stage and how has that been handled in previous analyses? In the case of the missing tumor stage, Stewart received his dataset and found that a large percentage of participants were missing this key variable. After checking the study documentation, he began asking study staff about this issue. Several conversations later, he learned that one data collection site of this multi-site study was unable to share tumor stage from patient medical records due to legal restrictions. Without this knowledge, Stewart may have been forced to delete these participants from his analyses under the assumption that their records were incomplete and unusable. When asked about documenting this discovery for other, he laughed and replied, *"No, it's more like an oral tradition" (Stewart, 352)*. Uncovering answers to such complicated questions could involve interactions with any number of people, ranging from just the local PI to collaborating study staff and PIs at other institutions. Post-docs described hours, even days, spent with data managers uncovering how variables were coded, as well as weeks emailing with the PIs at non-FHCRC study sites attempting to ascertain their study procedures and the resulting data. These efforts represent significant delays in scientific progress.

An Iterative and Ongoing Process

Sometimes understanding one aspect of the study simply opened the door to more questions by revealing the need for more information about other aspects. This happened in one of two ways. First, post-docs sometimes found they had used the data incorrectly or had used the wrong data altogether due to missing or incomplete information. Several participants noted that they had discovered problems with or gaps in their analyses while talking with their mentors and colleagues about their results. Kristina described presenting her analysis of the effects of dietary intake to her study group, only to find that she had used the wrong dataset. Because the datasets were not clearly labeled and several contained similarly-named variables, she had simply chosen the wrong one. Because no one in the study group could tell her the precise name of the correct dataset, she was forced to return to the study documentation and the files on the shared drive to try to identify the correct one by opening each one and comparing the variables inside to the ones she needed. This came after weeks of work on the incorrect dataset, significantly delaying her progress.

Ginger talked about how frustrating it was to go through this iterative process and how unproductive it made her feel, especially when compared to her previous work at another institution. She noted that having to constantly redo analyses based on the replacement of bad information with new information not only was a waste of time but made her look incompetent to her mentor. She stressed that better access to information could have prevented many of the issues she had experienced.

> *What happens is, I start off seeking the information, and I'm thinking about the study, so I have an idea of what information I wish I had, and then I have to do the analysis. And then I get a little bit of information, and so I say, 'I'm sorry; that analysis that I sent you was invalid,' and so then you look bad, which is just bad. And then you redo the analysis, and you're like, 'Oh no, in light of this new information it's apparently not appropriate to think of this way, and it would be better if this exclusion was made given the nature of the study sampling.'(Ginger, 177).*

The second possible cause for iterative information seeking was that the post-doc learned something scientifically interesting, either through their analyses or through conversations with colleagues, that required additional data be incorporated or explored. For example, learning that aspirin had a certain effect on the cancer of interest might compel the post-doc to look at the effect of other pain relievers or aspirin's effect on another cancer. As epidemiologists search for associations between variables and incidences of particular types of cancer, these additional factors (confounders and correlates) that are intertwined with, but not necessarily the cause of, the phenomenon under study must be addressed. What these factors are, precisely, can be different for different cancers or different populations and is based on deep understanding of both the dataset and current scientific knowledge. For example, smoking may have an effect on one type of cancer, but not another. Additionally, smoking can be a confounding variable for weight, because smokers tend to be of lower weight, making it difficult to know if the smoking or the weight is responsible for the association present in the data analysis. Irene talked about having her initial proposal and resulting manuscript reviewed by the study committee, who then recommended she consider incorporating other variables from the dataset into her analyses by addressing additional confounding variables, as well as exploring correlation between variables in the dataset.

> *So the first step is more of the content, or here's some other things you might want to look at. And also, if they have [done] that analysis on the same data set, they will tell you these are the useful variables, these variables might be confounding, so check for [those]. So those [are the] kinds of information I get and kind of, this variable is highly correlated with each other, don't use this one or [that] kind of information. But after the analysis I get lots of analysis comments as well, [about] the data and variables (Irene, 319).*

The first scenario was a frustrating setback and generally a waste of time, while the second was the result of interesting scientific work and moved the post-doc forward toward his or her goal. This second type of iteration is a normal result of a healthy scientific project, and we saw post-docs moving among the different question types noted in Table 1 throughout the entire data analysis lifecycle. In all cases, the post-docs once again returned to their information sources, both human and written, and sought the answers they needed to move forward with their analyses.

Understanding the Construction of Variables

As has been shown by previous CSCW literature, data are highly social, reflecting the values and practices of those who engage with the research and the data [5]. Data are also constructed as a result of the plethora of small, often seemingly insignificant, decisions and actions taken during a research project. However, knowledge of such decisions or actions often lives exclusively in the heads of those who took them, not through negligence but through pragmatism. Much like software engineers responsible for documenting changes in code specifications [13], it is simply impossible for researchers to document each decision formally while still moving their work forward. These decisions and actions may or may not be apparent in the data as irregularities or missing data. In either case, they can have a profound impact on the outcome of the data analyses if not understood, as they reflect the myriad of assumptions and interpretations, some explicit and some tacit, made by researchers and study staff along the way. In this section, we describe questions post-docs asked about the construction of variables as they worked to understand their data. These questions focused not just on the meaning of the variables, but on the decisions and actions that led to the current state of the dataset as presented to the post-doc.

How were these data constructed?

In order to use the data appropriately, post-docs needed to understand not just what a variable (e.g., "smoking status") meant, but how it had been constructed. In fact, the majority of difficult and time-consuming questions post-docs reported fell into this category. These were questions that focused on how specific variables were coded, how to match the questions on the original study questionnaires with the variables in the database and why the original study did what they did to the study data, including why they coded and analyzed variables the way they did. Such questions were often difficult to answer, as they involved digging deeply into analysis code, study documentation or people's memories, often across institutional borders. This process was rarely independent, but involved significant collaboration between the post-doc and the study staff.

Ginger struggled mightily with understanding how treatment variables in her dataset had been coded. She was interested in how a specific class of medications affected the risk of cancer in her population. According to the labels in the database, her variable described the duration of the medication use in months for a given individual. After completing her analyses, she worried that something wasn't quite right. Lengthy discussions with the data manager, which involved excavating five-year-old code and painstakingly reviewing each line in the coding of this variable, revealed inconsistencies between the label and the way the variable had been coded. This inconsistency had not been an issue in the original study because the variable was not a part of those analyses, as it was outside the study's specific aims.

> *I had received a variable for exposure to [medication] use that was months of [medication] use, and I assumed from the labeling of the variable in the dataset … that an individual was coded as a 1 if they had used at least one month of [medication], and so I said that our study excluded women who didn't have less than a month of [medication]. But apparently, women with one day of [medication] were counted as a one because it was really a zero to one month. And that took forever to figure out … because then [the data manager] had to go back to the original code in which she had created the variable and reinterpret the code to break down exactly what had happened, and it was like all this looping. So things like that were frustrating. We had a lot of setbacks where you're like, "So what does this variable really mean?" and every time you ask, "What does this variable really mean?" it's never straightforward (Ginger, 55).*

Had she simply accepted the label and not questioned it, Ginger's analyses would have been incorrect, and she would have made inaccurate claims in her paper. For any scientist, this is a terrifying prospect.

Some variables were difficult to understand because of the way the original question was asked. Stewart mentioned a variable whose question was vague, leading him to wonder how the variable had been interpreted in the coding scheme. In the excerpt below, Stewart wants to know if a given individual had a screening colonoscopy, as opposed to one that was diagnostic, a question that is difficult to answer based on the way the original question was asked. Yet understanding this distinction was crucial for his analysis. So he finds himself interpreting what the original study "thought they were asking."

> *It's also just sometimes there's nuances you want to understand. You know, someone's colonoscopy history, there's some questions about that, and you want to understand it, trying to rule out the colonoscopy that was used to diagnose the colorectal cancer. So you want to know if they were screened ever. And actually, that's kind of hard in this dataset because of the way the questions were asked, which is unfortunate. But so you're using like a time between diagnosis and when they say they had a colonoscopy to infer maybe what that was about. … And you have to go back to the data dictionary, for sure, but I find myself going back to the questionnaire to see what I think the question really was asking, you know. Because the data dictionary… it's just descriptive, it's kind of what they thought they were asking… it's like oh, the value can be one or two. One is yes and two is no. And, you know, it'll say had a colonoscopy [Laughter]. But when you look at the question, it's have you ever had a colonoscopy, you know, more than two years ago or something? So there's a difference. So there can be differences (Stewart, 162).*

Here, Stewart is questioning not just the actual meaning of the variable in the database, but how the question used to collect the data from research subjects was interpreted by researchers as the variable was constructed during the original study. In fact, some of the most difficult questions to answer in this category centered around the interpretation and decision-making history of a given variable. Post-docs noted that it wasn't simply a case of how the variable had been coded but why it had been coded a certain way. There is scientific knowledge encoded in these decisions, and post-docs were wary of proceeding with analyses that may have contradicted that previous knowledge or decision.

Stewart also mentioned that he wished more of the standard variables for his study, like pack-years of smoking, were derived by the study's central coordinating center, as opposed to being derived by each individual user with whom the dataset has been shared. He noted that variables were derived for previous papers in a certain way, and it was important to keep that consistent, yet those variables were each derived by individual investigators instead of by the central authority, leaving open the possibility for future mistakes by others. He felt it would reflect badly on the study and raise doubts about his analyses if he came up with a different result for, say, the average number of pack-years of smoking in the study than that which had been reported by earlier studies.

Reconstructing the history of a variable, then, is a complex, time-consuming and highly collaborative process. For their variables of interest, post-docs required a greater understanding of not only how the variable arrived at its current state, but also why. Various information sources were deployed to answer these questions; codebooks, questionnaires and the databases themselves were often used in tandem to ascertain an answer. Previous manuscripts were consulted to see how the data had been written about in the past, and to try to pin down how a derived variable may have been constructed. Data managers, original study staff and PIs were, however, the most frequently used information sources for finding

answers to these questions. There was simply no substitute for the institutional memory of those most intimately involved in the collection, coding and analysis of the data. Answering questions about the construction of the data often required substantial time and effort on the part of our participants and others. Amy discussed tracking down former study members in their new jobs, years later, in order to get answers to her questions. Without such information, post-docs did not feel comfortable moving forward with their analyses.

DISCUSSION & IMPLICATIONS FOR DESIGN

Our investigation of post-docs' use of preexisting data reveals new considerations for the support of collaboration around data reuse. While previous research has examined the difficulties of sharing data, focusing on assessing trust and reliability of the data, here we have focused on the issues post-docs faced while actually using data collected by others. They utilized a variety of strategies in their practice of science to gather enough information about the data that they felt comfortable using the data they had been given. Once we have moved past the issues of trust and reliability to focus on the actual use of preexisting data, questions of appropriate use become first and foremost.

It is clear that, even between trusted colleagues, sharing and reuse of data is still complex and fraught with pitfalls. That said, there was a logic in how post-docs went about answering the questions that kept them from completing their analyses. We have described here the different actions in which researchers engage to ensure proper usage of data, including nine questions to which post-docs needed answers as they moved their projects forward. One way to support better reuse of data is to provide better support for finding answers to this set of questions through better information management. Given that the majority of questions, and the most difficult and time-consuming questions, focused on the construction of specific variables, it would seem that documenting decision histories and variable coding procedures would be a good place to start. However, it is not practical to expect researchers to document a dataset sufficiently for all potential future uses. What we can do, though, is lend support to the collaborative process of information seeking by both helping studies organize their information better and aiding the conversations between the original study staff and the new data users.

The process of understanding data is not simply a matter of access to documentation of facts about or context of a study. Post-docs were not concerned only with details such as number of participants or what the variable "smoking status" meant, but with intricate decision histories of specific variables. As discussed above, however, completed studies and datasets are composed of a multitude of such decisions, some large and some small, some explicit and some tacit, all of them affecting the structure and scientific meaning of the data. For all practical purposes, it is simply not possible to document them all or to anticipate accurately which ones will be required by new investigators later. Without a mandate or cultural expectation to do so, it is also not realistic to expect that study personnel will spend significant amounts of time and money documenting a study not for their own purposes but for as-yet-unknown others. Our participants had direct access to all of the documentation available for a given study, as well as the original study personnel, and yet they still struggled with understanding certain aspects of their dataset. Furthermore, even if study personnel could somehow sufficiently document all of these decisions made about the data, we do not currently have a way to store such information that makes it easily accessible to data users.

In addition, post-docs were not concerned with an overall documentation of the context of a study, but with the documentation of the specific aspects of interest to their own, new analyses. The post-doc's areas of interest could be quite different than those of the original study staff. Much like a scientific database is designed to support specific research question [4], the information resources about a study are designed to support a specific project. When someone comes in to do new analyses on the data, especially if those new analyses are in a different discipline, they will require information not required by the original study.

For the post-docs in our study, the work of understanding a dataset and its variables is an ongoing and potentially unending process. Participants were working on complex, multi-institutional studies of hundreds or thousands of patients over long periods of time; there is always more to understand if one goes digging. Precisely which aspects of the study post-docs required understanding of depended entirely on their research question; they were concerned exclusively with those aspects that impacted their own projects. Information seeking stopped when they felt they had sufficient information to use the data correctly.

CONCLUSION

In this paper, we have explored the data reuse practices of post-doctoral researchers in cancer epidemiology as they worked to understand preexisting data for their own, independent analyses. We found that our participants had unmet information needs throughout the lifecycle of the research project, needs they addressed through a variety of strategies, including the use of both written and human information sources. This work of understanding the data was collaborative, iterative and ongoing, as new knowledge about one aspect of the data often sent post-docs back to their information sources to gain a deeper understanding of other aspects. Of special interest to our participants was the decision history that led to the construction of their specific variables of interest. They wanted to know not just how a variable was constructed, but why it was constructed the way it was.

As researchers are required to deposit their data into shared repositories, as funding agencies and open science advocates are demanding, they must do so with the assurance that others will be able to use the data in a way that is scientifically appropriate. Otherwise, the time spent documenting data for future use will simply be wasted effort.

This study has shown that even those investigators with direct access to all available study documentation and original study staff still struggled to make sense of the data they had been given. While enhanced documentation is a worthy goal, it is an elusive one when we consider the myriad of ways a dataset can be used by future investigators, ways which we cannot even imagine now. Supporting data reuse requires much more than just better documentation and creation of repositories. Rather, it requires the development of practices and processes that simultaneously support data sharing and reuse for the future while still supporting the current study. In other words, investigators need ways to engage in data curation in support of tomorrow's research without delaying today's. Finally, data reuse also would benefit greatly from funding agency incentives to data collectors to engage in data curation throughout the entire scientific process, as well as support for answering questions from users in the future.

The field of CSCW can play an important role in developing practices and systems that support the complex scientific collaborations that are now working to solve some of society's most pressing problem in energy, the environment, and health—including helping to treat and prevent cancer, as with our study participants. As scientific research continues to increase in both size and scope, data reuse, data sharing, and collaboration become increasingly important to the conduct of scientific research. Current practices in data sharing and reuse carry significant costs in time and effort, delaying science and diverting those resources from new discoveries. Systems and tools that could help staff document their data in such a way that users can find the right dataset the first time, or that could uncover the coding scheme behind variables without requiring weeks reading comments in analysis code, would go a long way toward creating greater efficiency and productivity in data reuse. Researchers need guidance on how to organize their information and to track those aspects of a study most likely to be questioned in the future; identifying those best practices will require more study on our part.

As more focus is turned to data reuse, we must continue to investigate ways to support more efficient data practices throughout the lifecycle of scientific projects. This includes a deeper understanding of how much information is enough to make a data user feel comfortable with their dataset, as well as what types of information best serve that need. Further, is it possible to develop training guidelines for both study personnel and new data users that highlight best practices in data reuse? Perhaps frustration levels can be minimized simply through better organization of information or the presentation of information at different stages of the project lifecycle. These are all questions for future research. Building on our knowledge of issues of trust and reliability, we have much to gain by developing an understanding of how scientists can use data collected by others in order to answer difficult scientific questions, allowing us to alleviate some of these difficulties and better support scientific innovation.

ACKNOWLEDGMENTS

This research was supported in part by the Fred Hutchinson Cancer Research Center and by NIH award R03CA150036. We would like to thank our participants for their generosity with their time and our anonymous reviewers for their thoughtful feedback. We also thank Dr. John D. Potter for sharing his expertise on epidemiology and Dr. Polly Newcomb for early discussions on post-docs and data sharing that sparked this research question.

REFERENCES

1. Baker, K.S., & Yarmey, L. (2009). Data stewardship: Environmental data curation and a web-of-repositories. International Journal of Digital Curation, 4(2), 1-12.
2. Bietz, M.J., Baumer, E.P.S., & Lee, C.P. (2010). Synergizing in Cyberinfrastructure Development. Computer Supported Cooperative Work, 19(3-4), 3-4.
3. Bietz, M.J., Ferro, T., & Lee, C.P. (2012). Sustaining the development of cyberinfrastructure: an organization adapting to change. In *Proc. CSCW 2012*, ACM Press (2012), 901-910.
4. Bietz, M.J., & Lee, C.P. (2009). Collaboration in Metagenomics: Sequence Databases and the Organization of Scientific Work. In *Proc. ECSCW 2009*, Springer-Verlag (2009), 243-262.
5. Birnholtz, J.P., & Bietz, M.J. (2003). Data at work: supporting sharing in science and engineering. In *Proc. ACM SIGGROUP*, ACM Press (2003), 339-348.
6. Borgman, C.L. (2011). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6), 1059-1078.
7. Edwards, P.N., Batcheller, A.L., Mayernik, M.S., Borgman, C.L., & Bowker, G.C. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, 41(5), 667-690.
8. Faniel, I.M., & Jacobsen, T.E. (2010). Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. Computer Supported Cooperative Work, 19(3-4), 355-375.
9. Fortier, I., Burton, P.R., Little, J., et al. (2010). Quality, quantity and harmony: The DataSHaPER approach to

integrating data across bioclinical studies. *International Journal of Epidemiology*, 39(5), 1383-1393.

10. Fortier, I., Doiron, D., Little, J., et al. (2011). Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. *International Journal of Epidemiology*, 40(5), 1314-1328.
11. Karasti, H., Baker, K., & Halkola, E. (2006). Enriching the Notion of Data Curation in E-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network. *Computer Supported Cooperative Work*, 15(4), 321-358.
12. Lee, C.P., Dourish, P., & Mark, G. (2006). The Human Infrastructure of Cyberinfrastructure. In *Proc. CSCW 2006*, ACM Press (2006), 483 - 492.
13. Lethbridge, T.C., Singer, J., & Forward, A. (2003). How Software Engineers Use Documentation: The State of the Practice. *IEEE Software*, 20(6).
14. National Postdoctoral Association. 2009. Accessed May 29, 2012. <http://www.nationalpostdoc.org/policy/what-is-a-postdoc>
15. National Institutes of Health Research Portfolio Online Reporting Tools (RePORT). 2012. Accessed May 29, 2012. <http://report.nih.gov/award/index.cfm?ot=&fy=2011&state=&ic=&fm=&orgid=#tab2>
16. Nelson, B. (2009). Data sharing: Empty archives. *Nature*, 461(7261).
17. Piwowar, H.A. (2011). Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data. *PloS one*, 6(7), e18657.
18. Piwowar, H.A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PloS one*, 2(3).
19. Walport, M., & Brest, P. (2011). Sharing research data to improve public health. *The Lancet*, 6736(10), 9-11.
20. Zimmerman, A. (2007). Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries*, 7(1-2), 5-16.