Data Reuse and Sensemaking among Novice Social Scientists

Ixchel M. Faniel
OCLC Research
6565 Kilgour Place
Dublin, OH 43017-3395
fanieli@oclc.org

Adam Kriesberg
University of Michigan
School of Information
105 S. State Street
Ann Arbor, MI 48019-1285
akriesbe@umich.edu

Elizabeth Yakel
University of Michigan
School of Information
105 S. State Street
Ann Arbor, MI 48019-1285
yakel@umich.edu

ABSTRACT

We know little about the data reuse practices of novice data users. Yet large scale data reuse over the long term depends in part on uptake from early career researchers. This paper examines 22 novice social science researchers and how they make sense of social science data. Novices are particularly interested in understanding how data: 1) are transformed from qualitative to quantitative data, 2) capture concepts not well-established in the literature, and 3) can be matched and merged across multiple datasets. We discuss how novice data users make sense of data in these three circumstances. We find that novices seek to understand the data producer's rationale for methodological procedures and measurement choices, which is broadly similar to researchers in other scientific communities. However we also find that they not only reflect on whether they can trust the data producers' decisions, but also seek guidance from members of their disciplinary community. Specifically, novice social science researchers are heavily influenced by more experienced social science researchers when it comes to discovering, evaluating, and justifying their reuse of other's data.

Keywords

Communities of practice, Data repositories, Data reuse

INTRODUCTION

Scholars have examined data reuse within and across disciplinary communities to determine whether and how people make sense of each other's data (e.g., Birnholtz & Bietz, 2003; Carlson & Anderson, 2007; Schiff, Van House, & Butler, 1997; Van House, 2002; Zimmerman, 2007). One of their major goals has been to inform the development of a supportive data repository environment. Although data can be disentangled from their context of production, the work of data scientists and digital archivists increases as the data circulate farther from their point of origin (Berg & Goorman, 1999). Disseminating data more widely within and across disciplines impacts the amount and kind of context that needs to be captured and communicated (e.g., Jirkota et al., 2005; Karasti, Baker, & Halkola, 2006; Wallis, Milojevic, Borgman, & Sandoval, 2006). The amount of additional work to process data is evident in the transition from research collections to resource and reference collections (Baker & Yarmey 2008; Cragin & Shankar 2006; National Science Board 2005). As the data move further away from their native environment supporting and supported by the research team that produced the data and are more widely disseminated to single or multiple disciplinary communities, the data's context becomes more important as does knowledge of the needs of the data users. For instance, large-scale data reuse over the long term depends in part on uptake from early career researchers within the discipline who are still learning the norms and conventions of their communities' research practices. Yet, we know little about how data should be disseminated to novices for successful reuse. Knowing whether differences between novice and expert data users exist is important for data repositories. This is particularly true for those seeking the Trustworthy Repository Audit and Certification (TRAC) or the Data Seal of Approval and must demonstrate an understanding of their designated communities. To support these aims, we begin a line of inquiry to examine novice social science researchers' (NSSRs) reuse practices. For this study, we ask: *How do NSSRs make sense of social science data?* A follow-up study to examine more experienced social science researchers (ESSRs) is currently underway.

ASIST 2012, October 26-31, 2012, Baltimore, MD, USA.

LITERATURE REVIEW

Much of the data reuse literature has drawn from the concept of communities of practice to explain reuse behavior within and across disciplinary communities. The research suggests that data reuse is easier when data circulate within as opposed to outside of a disciplinary community, because members of a disciplinary community

share understanding, practices, technology, artifacts, and language (e.g. Jirkota et al., 2005; Van House, Butler, & Schiff, 1998). For instance, formal and informal training and data collection experience in their discipline helped ecologists select and evaluate others' data for reuse (Zimmerman, 2007, 2008). Similarly, earthquake engineering researchers relied on their familiarity with the content and structure of journal articles and data documentation to select and understand data produced during complex laboratory experiments (Faniel & Jacobsen, 2010). However, these skills, knowledge, and experiences develop over time. Only as new members of a community master the requisite skills and knowledge through a process of legitimate peripheral participation do they move toward full participation in the sociocultural practices of their community (Lave & Wenger, 1991). Drawing from the communities of practice literature, we suspect that novice data users learn how to reuse others' data from old-timers (i.e. experts). We also suspect that novices, on their way to becoming experts, behave differently. Yet we know little about whether and how these differences might influence their data reuse practices, particularly how they make sense of others' data.

Studies examining novice data reuse practices have focused on reuse outside of one's own disciplinary community (e.g., Wallis et al., 2006; Weeks & Packard, 1997). Reuse outside of one's discipline is challenging because differences between the disciplines make it difficult to communicate information across them (Palmer 1996; Pierce 1999). These differences include the expectations of those involved in the peer review process, the models or paradigms on which research is based, and the distinct stylistic and presentational features that exist in each field (Pierce, 1999). For instance, ranchers and oyster fishermen reusing scientific data understand and judge scientific claims according to how well the scientific explanations match their personal experiences or their community's historical relationship with the scientist's institution (Weeks & Packard, 1997). This view contrasts that of scientists, who evaluate the data based on their understanding of the methodological approach and its rigor (Weeks & Packard, 1997). Several studies have shown how data repository staff have addressed these differences. Staff at the Center of Embedded Network Sensing (CENS) provide special tools and filters as scaffolding for high school science students whose skills and epistemologies differ from the ecologists and environmental engineering researchers who produce the data (Wallis et al., 2006). In addition, CalFlora staff have changed procedures to accept plant distribution data from institutions and personal sources, including the general public (Van House, 2002). By adding metadata, such as the contributor's level of expertise and certainty of taxon identification, data users are expected to be better able to assess the credibility of contributions from others who are not members of their community (Van House, 2002).

Research examining how individuals reuse data from others outside of their disciplinary community has informed the development of socio-technical solutions to facilitate novices' data reuse. However the individuals examined are not fully representative of all novice data users. As more experienced members of other communities, the novices studied often have different skills, knowledge, and experiences than the community where they are getting the data. They also have had few if any opportunities to be acculturated into that community, especially when it comes to disciplinary-specific practices. We contend that more can be learned by examining novice users reusing data from others within their disciplinary communities. Therefore, in this study we examine early career social science researchers (e.g. graduate students, post docs, assistant professors) who have worked on at least one project involving data reuse, but have not published more than one paper. They are novices who have less experience not only reusing data, but also conducting social science research. The next section of this paper will detail the research methodology, including the data collection site, and our data collection and analysis procedures.

RESEARCH METHODOLOGY

The Dissemination Information Packages for Information Reuse (DIPIR) project is studying three diverse disciplinary communities: zoology, archaeology, and quantitative social science (Faniel & Yakel, 2011). The current paper focuses on Inter-university Consortium for Political and Social Research (ICPSR), a data archive for the social sciences located at the University of Michigan in Ann Arbor, MI.

Site Description

ICPSR was established in 1962, and currently has 116 permanent staff members. As one of the largest repositories of its kind, it houses 50,000 files of primarily well-structured, digital, quantitative data. ICPSR adheres to best practices in digital preservation and data archiving. It is an OAIS-compliant repository and has been awarded the Data Seal of Approval. Data enters ICPSR through several means. Any producer of social science data can choose to submit a dataset via ICPSR's online submission form. In some fields, data deposit with ICPSR is mandated by funders. ICPSR also recruits data from major studies and contracts with several survey organizations and federal agencies to archive their data.

ICPSR's designated community of users consists primarily of researchers at its 700 member organizations, which include colleges, universities and research centers in the U.S. and abroad. They are quantitative social scientists working in a variety of fields, some of which align directly with ICPSR's 16 topical archives on demography, aging, substance abuse, and mental health, among other areas. While most of ICPSR's data holdings are available to the general public, access to other datasets is restricted to researchers who are affiliated with member institutions.

Data Collection

We conducted 22 semi-structured interviews with NSSRs asking respondents to discuss their experiences reusing quantitative social science data. Topics of inquiry included how respondents discovered and evaluated data for reuse and their experiences with ICPSR and other data repositories. Interviews were approximately 1 hour and respondents were paid $25 for their participation in the study. All interviews were recorded and transcribed.

Data Analysis

The interviews were analyzed using NVivo, a qualitative data analysis software tool. Our team worked collaboratively to develop the code set and two team members coded the interview transcripts. The code set was developed based on the interview protocol and expanded as new codes arose from our ongoing analysis. For instance transcripts were coded for mentions of how NSSRs discovered and selected data and their purpose of reuse. The transcripts were also coded for different aspects of the context of data production NSSRs mentioned, such as research design, associated documentation, and data descriptive information to name a few. After training on several test transcripts, the coders reached a reliability of 0.8 using Scott's Pi, a statistic showing high inter-rater reliability for the coding of textual data.

RESULTS

It was not surprising that our analysis of interviews showed NSSRs who primarily dealt with quantitative social science data were interested in making sense of the numbers. As one respondent explained, the issue was not just that the data were numerical. It was also that the numbers often represented concepts that could be defined and measured in different ways.

…it's numerical value on things that don't have numerical value. So it's not like a sort of thing is worth a certain amount, that numerical value is something that everybody can understand (CBU14).

More specifically, we found NSSRs were particularly interested in understanding how data

1. were transformed from qualitative data to quantitative data,
2. captured concepts not well-established in literature, and
3. could be matched and merged across multiple datasets

In the paragraphs that follow we discuss how NSSRs made sense of the data under these three circumstances: transforming qualitative data to quantitative data, capturing concepts not well-established in the literature, and matching and merging data across multiple datasets. Our findings showed NSSRs wanted access to enough of the context of data production to make sense of data producers' rationale for methodological procedures and measurement choices. Part of their process included a carefully detailed articulation of the data producers' research processes during which time they asked themselves whether they believed the data producers made the right decisions. NSSRs also sought out ESSRs in the disciplinary community to help manage complex reuse decisions and processes.

Transforming Qualitative Data to Quantitative Data

The NSSRs' careful articulation of the data producers' research processes were evident in the transformation from qualitative data to quantitative data. Stepping through detailed data collection and coding procedures helped them develop trust in the data and in their results. Much of the quantitative social science data that respondents described was produced from qualitative data. For instance, NSSRs spoke of nominal data created when a categorical label was transformed into a number for use in a statistical analysis. A common example was transforming the gender labels male and female to 1 and 0, respectively. Transforming (i.e. coding) the data typically resulted in a direct mapping of qualitative data to nominal data (e.g. male=1, female=0) and appeared in documentation (i.e. codebooks) describing the data.

Similar to gender, race was another categorical label that was often transformed. Although direct mappings were also used for race, it was not enough for one NSSR interested in studying African American middle school students. CBU10 also needed details about the data collection instruments and procedures to address questions she had about how the parents and students were allowed to identify themselves.

Since I'm looking at African-American middle school students, I want to find out when they ask the question to the parent or to the student, how was that question asked and was there follow-up questions in terms of did they ask what is your race as opposed to allowing the parent or the student to tell them what their race was. And if they did ask that question, did they also follow up to find out whether or not race was used as a specific category or were there other categories within that particular indicator that would allow parents or the student to determine was that a full... Did that represent their total identity or were there other follow-up questions that would allow them to determine mixed identities (CBU10).

Although the differences in data collection CBU10 described were subtle, they had a huge impact on her research and her confidence that the data were reliable. She tried to get at the subtleties by examining the data collection procedures. She not only wanted to know what questions were asked during data collection, but also how the questions were asked, and what were the follow-up questions, if any. Following advice from her econometrics professor she described why looking at the data producer was only part of the process.

[J]ust because you get a dataset doesn't necessarily mean that the dataset is a reliable dataset. And so, there are

different statistical tests that you can use in order to determine the results…looking at who actually collected the data, that's just a part of the process. Look at the total process and determine, try to find out as much information as you can (CBU10).

Asking detailed questions and gathering information to describe how the data producer conducted the research coupled with statistical analysis helped her evaluate the reliability of the dataset.

For other NSSRs, knowing how the mapping was developed (i.e. coding procedures) to transform qualitative data to quantitative data was important. CBU12 was interested in the relationship between legislative hearings and news coverage and wanted to determine which one had more influence on a state's agenda. Reading about the coding procedures data producers used when reading and assigning numerical values to the newspaper articles in particular allowed CBU12 to see and really understand what those data represented.

And they were telling them how each [newspaper] article should be coded based on what it was describing. It's been a while since I went through the book. But that's kind of it. I mean, the basic gist is you have this article. These are the different things that you need to look for in this article, to assign a value to it. Those values came from those policy codes, and also whether it was a House or a Senate hearing, they had to copy down the headline of the article and then, like the first paragraph or the summary of it or something like that. They were just explaining all those pieces that were necessary to try to keep the reliability among the different universities… (CBU12).

Similar to CBU10, talking through how the data were constructed rather than just trusting that *"…everything is okay"* was an important part of CBU12's reuse experience. However he also noted that such details were not always provided with the data.

Developing trust by understanding and explaining someone else's data was particularly important when NSSRs tested their theoretical models. It meant NSSRs could trust their own results. As CBU03 explained, getting non-significant results when reusing others' data would cause him to question the data. Using the New York Times as an example, he explained the consequences of not knowing the coding procedures used to develop a ranking representing the "democratic-ness" of countries based on newspaper articles.

So they use New York Times continuously for the 30 years. New York Times, it has changed. So I want to know what years New York Times was used to gather data. I'm sure they used more than one newspaper. Also, I want to know which ones those were, for example. It might be an example, only because of reliability again. I want to … in case I need to refer back to the data and look at …, I am having some complications, if I really trust my theory and the stats is saying something else maybe because of an outlier or something like that, I want to be able to go back and determine if there's something that could have been... If there's a reason why my theory isn't aligning with my results or something like that (CBU03).

CBU03 acknowledged that it would be a long shot if the data were the cause of his non-significant results; however, it was something that he wanted to be able to rule out. CBU03 believed knowing more about the data production process, particularly what years of the New York Times were used to code for "democratic-ness" and what specific papers within those years, would answer his questions.

Prior to data reuse, CBU03 was driven to want to know "*as much possible about how they [the data producers] gathered the data…As well as...in the mind of those gathering the data, what they had in mind basically"*. He explained that as a new scholar, he was constantly asking questions about whether he was operationalizing things correctly. Walking through how data producers' did it for their research studies helped. For CBU03 this meant reading data producers' justifications for their research decisions in the codebooks.

In sum, NSSRs often sought much more than the simple maps showing the transformation of qualitative labels to numerical data. Subtleties in the data collection and coding procedures helped them to understand how the maps were created and to determine data reliability. Their step by step construction of the data production process helped to determine whether they trusted the data producers' data, which in turn increased trust in their results from data reuse.

Capturing Concepts Not Well-established in Literature

When confronted with reusing data that did not have well established measures, again we found respondents stepped through the data collection methodology in detail to understand how those data were produced. However we also found they also sought guidance from the social science community through publications that critiqued and supported the data.

A key concern when reusing data that lacked agreed upon measures was the possibility of bias in the data collection methods, because there were no agreed upon definitions or measures. In the following quote, CBU09 described concerns with human rights pressure data she was interested in reusing.

If I were to be looking at a dataset from like... That had been collected by a researcher for the [Institute Name] that would be something that I'd want to keep in mind considering that it is a group that is really trying to advocate for a particular direction in foreign policy. And that's not to exclude it just by the nature of it being a right wing organization, but I would want to evaluate their methods to see if that's the methods that I would've chosen and see if those methods... I mean, I think everyone knows

that when it comes to data, there's a lot of ways to play around with it to make your point (CBU09).

In another example, CBU13 explained how federalism had been defined in many ways and people had employed different rationales to decide whether to code countries as federal or not. Thus the question for CBU13 was not only did the data meet her research objectives, but also was the way federalism was defined and measured match her personal beliefs. In her case, going back to the journal article that described the data producer's definition and coding procedures for federalism helped her answer those questions.

…there are many definitions of federalism out there. There are not a lot of federal systems, but like some people put one country as federal and others don't put it as federal. Some people put China as federal and some other people say it's not federal. So, I think especially with institutions it goes a lot to the conceptual definition. I think the conceptual definition and the actual coding process. So, what kind of coding decisions did this person make? Did he or she make a lot of what seems arbitrary decision[s]? Just trying to think about what my idea or my argument says and what I think conceptually of this process and see if I had done the coding, would I agree with that? Does that fit with what I believe (CBU13)?

Encountering data that do not have well-established measures, CBU09 and CBU13 were both concerned about the data's limitations. They wanted to ensure that the data collection methods were not biased and the data producer's thoughts behind conceptualizing and operationalizing the data were well formed and fit their personal beliefs. They reviewed data producers' conceptual definitions, data collection and coding procedures, and available publications to more fully understand the data.

We also found NSSRs sought guidance from the social science community through publications about the data or definitions and measures of the concept the data was intended to capture. The community had a powerful influence when definitions and measures for concepts were well-accepted. In her dissertation work CBU13 explained how she was examining anti-corruption agencies and how they interact with the institutional environment. Even though there was an anti-corruption dataset available, CBU13 was leaning toward collecting her own data, in part *"because it [the anti-corruption dataset] doesn't take the mainstream definition from literature for these agencies"*.

In contrast, when the community did not have well-established definitions and measures, NSSRs had more freedom to use the literature to make a case to reuse or not. For instance, CBU14 rationalized the decision to reuse regime type data by explaining that data were always being critiqued and she had found several well regarded journal articles that tested regime type using the same data.

I mean even though they were critiquing its limitations, it still didn't mean that we couldn't use it or that it wasn't worth using. So a lot of them were saying that it fails to take into account this and that or it's too simplistic or all the concerns you'd normally get with any kind of indicator. I mean even with the GDP [Gross Domestic Product data], people say it doesn't capture, I don't know, work that is done that doesn't have market value. That's not right because it's also... People critique all these indicators all the time, it's just a matter of what you need to find and what you have available. So given that there is no real alternative to using that, we have to pick the ones that exist (CBU14).

CBU14 also observed that critiques tended to hold less weight for data in short supply. In the end, she decided to reuse regime type data from more than one repository explaining that data from two different sources exhibiting similar patterns would lend weight to her findings.

In another example, CBU17 discussed how his concern with reusing expert survey data on European integration stemmed from the number of experts rating political parties' attitudes toward European integration.

[One] of the things that I was immediately worried about is that some parties, [e]specially the smaller parties, had only like one or two experts rating them, in the Dutch case, which makes it not super reliable, so that's what's kind of like [it made me think,] "Oh wow, I need to be careful about that because that's..." So I was like, "Oh I should really pay attention that that's not going to hurt me in the long..." (CBU17).

Similar to CBU14, he turned to the literature for guidance from the community. After reading a methodological article that addressed concerns other community members had expressed about the data and finding published articles by others who also had reused the data, he decided to reuse as well. Several NSSRs spoke of wanting to understand both sides of the debate their community was having about the merits of data they were interested in reusing. Given this circumstance, they had the freedom to choose, which side was more closely aligned with their beliefs.

In sum, NSSRs used codebooks and data producer publications when deciding whether the conceptualization and operationalization of data measures not well-established in the literature were sound. However, when it came to addressing the impact the data might have on their research, NSSRs were guided by a bibliography of literature related to the data. They sought articles critiquing and supporting the data to build their awareness of the issues and merits of the data, such that they could then defend their reuse choice.

Matching and Merging Data across Multiple Datasets

NSSRs often consulted with faculty advisers about matching and merging data across multiple datasets, because it often involved more complex reuse decisions and

processes. Faculty advisors helped NSSRs make and provide the supporting rationale for reuse decisions as well as access data that was more appropriate for their research objectives. For instance, some of the research questions NSSRs were interested in required them to reuse data from a longitudinal survey. In other instances, NSSRs were combining data from multiple sources (e.g. repositories). In both of these situations they discussed dealing with challenges such as missing data, changes to data collection instruments, data from different time periods or at different levels of aggregation, and the need the identify or create unique identifiers. For instance, a unique identifier was needed for all the datasets to ensure that NSSRs were matching and merging data from the different datasets correctly. CBU07 explained how she used data about the states to combine data from multiple datasets.

So there's always one thing that you match on when you do the merging of the datasets. So if I were given clear information on the availability of this, for example if you were merging on the States, if you were merging and you needed... So you would need to know that the code variable for the State is on every dataset that you're going to use, right? But then that's always that you usually find out by looking at the codebooks of your different datasets you are going to merge. So that's usually the trouble that you run into that if it's not there you have to create it yourself and then create a variable but that's usually what you need or that's what I would be looking for (CBU07).

Similarly, CBU18 looked to match political identity data over time. A common challenge NSSRs confronted when dealing with longitudinal survey data however, were changes to survey questions and response scales over time as data producers experimented with wording. Sometimes the questions were dropped entirely. These changes affected NSSRs' ability to merge data across the different datasets. For some it also meant that they did not have a sample size large enough to conduct their intended statistical tests.

…so when I'm trying to get it at a complex variable like political identity. If they're not asking the same question over years, in my original development of the variables [it's] particularly difficult because if they've changed the question wording, are then people answering differently and so there were several discussions that I had with my dissertation advisor about, "Can we still use these if they changed the question wording?" "Is it different or the same enough such that it either should be thrown out or it should stay in?" So it is mainly about, again, constructing that important variable that I cared about (CBU18).

To deal with these types of challenges, CBU18's had conversations with her dissertation advisor to better understand how changes to the survey questions measuring political identity impacted her dissertation research. She also mentioned additional contextual elements associated with data production that would have helped her make sense of the changes that were not made available. For instance, simple things, like knowing why questions were put in and taken out of longitudinal survey, would have answered some of her questions, such as: "*Is it not measuring what they wanted it to measure? Did they find it wasn't valid variable or where too much variance? What happened there? What's the data piece of why they took it out?*" (CBU18).

Wanting to reuse 4-5 years of media data from a European survey, CBU02 faced similar issues as CBU18. However, in contrast to CBU18, he preferred not to use the data, in part because he felt some of the changes did not reflect accurate measurement of the phenomenon.

…sometimes they change a little bit in terms of just asking people how they just get their news, or how much TV they watched, how many newspapers they read. [S]o they might ask, on a typical week how many times a day do you, let's say, watch TV? Let's say you would say one to seven. Or they might ask how often do you watch TV, not at all, regularly to very frequent. [S]o the sort of not at all, regularly, frequently is I'd say it's not very accurate. So I prefer not to use those ... that type of data (CBU02).

In addition to combining data from a longitudinal survey into a time series, respondents described merging data from different sources. For instance, CBU04 combined data from two repositories. Although cautious when matching datasets from different sources, he liked to rely on the literature as well as rationale provided by his professors to justify his decisions if possible.

I guess I'd just be more cautious just to make sure that I'm matching up the right time frames mostly. It tends to be… At least for me, it's an issue of making sure the dates are correctly merging with the information I have involved. So, a lot of times, authors will create a variable, they'll average across a four or five year period, and I'm trying to match that with a variable that was coded for a single year period. So making an argument and suggest[ing] that these two things should be put together and measured in the same time period, is something I always have to be wary of and cautious of. So when dealing with that, it's just that I've generally been either an appeal to authority, so I'll see if it's been done by others. If I can't make sense of it to myself and then I'll ask professors to give me some rationale for why I'm doing what I'm doing. But in general, it tends to be a good reason for doing it, I think. Convinced myself there's a good reason to do it… (CBU04).

We found the aggregation process could become quite complex depending on the number of different datasets, data collection methodologies, and units of analysis involved. Examining capital punishment and geographical disparity in death sentencing required merging data from several sources. Choosing county as the unique identifier for each dataset, CBU05 explained how the Supplemental Homicide Report (SHR) data had to be aggregated from agency to county levels first. This meant using additional

datasets not for the data per se, but for crosswalks that made the aggregation possible.

Some of it was ICPSR data, I used the SHR for my control variable and LEOKA [a report on law enforcement officers killed in the line of duty] for my control variable, and then I used the Uniform Crime Report for the violent crime rate. And then I also used a few other datasets through ICPSR that weren't so much for the data in particular for a variable but like crosswalks and things like that they make available. Like for matching some of the data, like the SHR data, it's all organized by ORIs, the codes for the actual police agencies so to put it at the county level ICPSR has a dataset available with all of those organized by county (CBU05).

CBU05 also explained the help she got from one of her instructors. The instructor had not helped decide whether or how to use the data, but rather assisted her in securing Law Enforcement Officers Killed and Assaulted (LEOKA) data at the individual rather than the aggregate level, because it was more in line with her research objectives.

Like the LEOKA data that they make available through the archives is all aggregate so you can't ... but you can tell from the reports that are released each year, they actually have that information in the reports, not on the individual level, but they say, this percentage of officers were white. So, I knew the data had to be out there somewhere and she [my instructor] was very helpful with contacts in the FBI to be able to find somebody who might have that data (CBU05).

Faculty also made NSSRs aware of particular datasets that were well-suited for their specific research interests. For instance, a faculty advisor pointed CBU15 to the National Corrections Reporting Program (NCRP) data after he expressed interested in conducting imprisonment research at local rather than state levels. Reusing these data with census data from the Integrated Public Use Micro Sample (IPUMS) at the Minnesota Population Center, CBU15 was able to engage in a complex aggregating and merging process to get data that were closer to meeting his research objectives.

So the files from NCRP are at the county level. IPUMS is what's called the public use micro area levels. So the... It's the smallest unit of geography that's publicly available is that the public micro area. And it's a collection of counties of a hundred thousand or more. So it could be one county or it could be a multitude of counties. So within that, so the census gives files, what they call boundary files. So okay, some of those areas change over time, like say, Ann Arbor expands and takes on more counties like there could be more counties in the year 2000 than the year 1980 and so on and so forth. So they have files that account for that in those areas exactly as they're defined. How that's linked, so the corrections data are each person is part of the information that's applied for each person is a county of sentencing variable. So within that is the five-digit code that identifies the state in the three-digit area. So I used the census boundary files to link people within those units to that bigger micro area and then I aggregate it. And I say, "Okay, of all these people, how many are in this county?" And then count which counties fit into this area (CBU15).

In sum, NSSRs faced with a range of issues when matching and merging data, including changes to data collection instruments and differences in data collection timeframes and units of analysis. In working through the issues, NSSRs often received help from faculty advisors to understand the impact the reuse decisions would have on their research, get supporting rationale for their reuse decision, and access data that were more appropriate for their research objectives.

DISCUSSION

We examined NSSRs who reused data within their disciplinary communities to offer an alternative view on novice data reuse practice. NSSRs were particularly interested in making sense of how data 1) were transformed from qualitative to quantitative, 2) captured concepts not well-established in literature, and 3) could be matched and merged across multiple datasets. Our findings suggest that there are similarities and differences between NSSRs and other novice data users examined in prior research. For instance, when deciding to reuse the data, our findings indicated that NSSRs engaged in a more careful articulation of the data producers' research process. Our findings also indicated that NSSRs required scaffolding during their data reuse experiences, which is similar to prior research. The kind of scaffolding is different however, in that it took human rather than technological forms and helped manage rather than reduce the complexity.

Careful Articulation of the Data Producer's Research Process

As previously noted, prior research has focused on novices reusing data outside of their disciplinary communities. For instance, a study of ranchers and oyster fishermen reusing scientific data found that they rely on factors external to the research process to understand and judge the soundness of the data, such as their personal experiences and their communities' knowledge of the scientist and its historical relationship with the scientist's institution (Weeks & Packard 1997). In contrast, NSSRs engaged in a careful articulation of the data producers' research processes. Unlike the ranchers and oyster fishermen, they wanted fine-grained details about the data collection and coding procedures. For instance they were interested in finding out what questions data producers asked during data collection, how the questions were asked, and how a concept that the data was meant to capture was defined. When the data being coded came from secondary sources, NSSRs wanted to know details about those sources, such as the year of the source or the specific content examined for coding. NSSR's were very attuned to the research process.

NSSRs objective in getting their hands on such detail was to step through how the data were produced in order to determine the soundness of the data producer's methods to decide if they agreed with the approach, and to determine how to most appropriately reuse the data. In this regard, the NSSRs' actions resembled those of expert data users seeking data from colleagues in their disciplinary communities. For instance ecologists used context to visualize their colleagues' data collection process (Zimmerman 2008), while earthquake engineering researchers used it to understand colleagues' data as though they were present during data collection (Faniel & Jacobsen 2010).

Human Scaffolding to Help Manage Complexity

Similar to prior research, we found NSSRs required scaffolding, particularly when their decision to reuse or the process of reuse was complex. In prior research scaffolding has consisted of changes to the user interface to simplify scientific tasks (Wallis et al., 2006). In contrast, we observed NSSRs' using scaffolding in human form. They turned to members in the community to manage rather than simplify the complexity. Specifically, we found two types of human scaffolding, the community as represented in the literature and interactions with faculty advisors. The literature was used to make sense of data that captured concepts not well-established. Written for and by the community, the literature served as a forum for debate. NSSRs sought articles critiquing and supporting the data. Since the data represented concepts not well-established, NSSRs used the literature to understand all sides of an argument in order to begin to formulate their own views about reusing the data. Since there was freedom to choose sides, NSSRs also considered whether the data producer's approach fit with their beliefs. Looking at the data producers' publications as well as the codebooks, NSSRs asked themselves whether they would have taken the same actions as the data producer. Interestingly, when deciding what side of the debate to take, NSSRs also took notice of the reputation of the journal and author of the article as well as the arguments.

In contrast to the literature which acted as a forum for debate NSSRs could assemble, NSSRs' use of faculty advisors more closely reflected the traditional master-apprentice relationship. NSSRs often sought out or were given counsel by faculty advisors about how to proceed with reuse whether this involved merging multiple data sources capturing the same concept over time or matching data representing different time periods. Faculty advisors helped NSSRs understand the impact changes to data collection instruments had on their reuse of data. Faculty also helped NSSRs understand how these changes impacted measurement of the concept, how concerned NSSRs should be about certain changes, and how NSSRs should rationalize reusing the data in light of the impact. Faculty also helped NSSRs access data that was more appropriate for their research objectives by pointing them to alternative repositories or introducing them to personal contacts.

Future Research

One of the motivations behind this research study was to offer an alternative perspective on novice data reuse. Specifically, we were interested in understanding how novices reused data within their disciplinary community. Much of the prior research has examined people reusing data outside of their community and considered them to be novice users. While we do not want to argue who constitutes a novice, we do believe future research should begin to parse novice users in terms of those that are within and outside of the discipline from those that are less versus more experienced data users, and those that are less versus more experienced researchers as the need arises.

In our study, NSSRs were able to seek or were given faculty counsel. However this may not be the case in disciplinary communities where data sharing and reuse are new phenomena. Future research should examine what happens when novice data users do not have the support and guidance from more experienced community members. For instance, how are data reuse practices learned and passed to novices? What kinds of scaffolding can be usefully employed to support complex reuse decisions and processes? What is the responsibility of the advisor and what could repositories do to create scaffolding?

As we continue to analyze our NSSR data in conjunction with our study of ESSRs, we will be able to offer comparisons of reuse behaviors. Although we suspect that NSSRs and ESSRs need the same amount and level of contextual information, we are looking forward to seeing whether our analysis bears that out. Analyzing the NSSR and ESSR data together, our aim is to provide a much more detailed analysis of the context necessary for reuse for the two types of users. We are also interested in determining whether ESSRs have cause to use scaffolding. While keeping current with the thoughts and opinions of community members through the literature seems necessary, relying on advisors to help manage complex reuse decisions and processes seems less likely.

Practical Implications

Given NSSR reuse behavior closely resembles prior research on more experienced scientists reusing data from within their own communities, it seems less likely that data repository staff have to do additional work to capture different amounts and kinds of the context for novice vs. expert data users. However, they may have to do additional work to present the context they have in different ways until novices gain the requisite skills, knowledge, and experiences to become full participants in their disciplinary communities. For instance, data repositories should consider how to replicate the human scaffolding NSSRs employed.

Being able to poll the community through the literature particularly for data that capture concepts that are not well-established was important. ICPSR provides an online bibliography of data-related literature. However, our sense from NSSRs is that access to bibliographies in general can be uneven – i.e. only listing supporting literature and/or the data producer's publications. As our findings showed, accessing critiques and support for data reuse in the literature as well as data producer publications are critical.

Faculty advisors are more difficult for repositories to supply; however, it will be especially important for disciplinary communities where data sharing and reuse are new and faculty advisors are less plentiful. One of the ways ICPSR addresses this issue is through its Summer Program in Quantitative Methods of Social Research. It offers several courses each year that introduce participants to particularly complex datasets and helps them mange complex reuse decisions and processes.

CONCLUSION

To be successful, data repository staff must know and support the needs of their designated community of users. Distinctions have been made between expert and novice data users, but they tend to focus on novices reusing data outside rather than within their own disciplinary community. Although we found NSSRs' reuse behavior resembled that of experienced data users, we also found they used members of their community as scaffolding when reuse decisions and processes were complex. Understanding how NSSRs engage with ESSRs and how that influences their access to and reuse of the data stored in a repository can be useful for data repository staff to monitor. The results of such monitoring could influence how data and associated context are presented and what scaffolding tools and services should be provided for novices on their way to becoming experts.

ACKNOWLEDGMENTS

This research was made possible by a National Leadership Grant from the Institute for Museum and Library Services, LG-06-10-0140-10, "Dissemination Information Packages for Information Reuse".

REFERENCES

Baker, K. S., & Yarmey, L. (2008). Data Stewardship: Environmental Data Curation and a Web-of-Repositories. *4th International Digital Curation Conference*. Edinburgh, Scotland.

Berg, M., & Goorman, E. (1999). The contextual nature of medical information. International Journal of Medical Informatics, 56(1–3), 51–60.

Birnholtz, J. P., & Bietz, M. (2003). Data at Work: Supporting Sharing in Science and Engineering. *ACM Conference on Supporting Group Work* (pp. 339–348). Sanibel Island, FL.

Carlson, S., & Anderson, B. (2007). What Are Data? The Many Kinds of Data and Their Implications for Data Re-Use. *Journal of Computer-Mediated Communication*, *12*(2), 635–651.

Cragin, M. H., & Shankar, K. (2006). Scientific Data Collections and Distributed Collective Practice. *Computer Supported Cooperative Work*, *15*(2-3), 185–204.

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*, 355–375.

Faniel, I., & Yakel, E. (2011). Significant properties as contextual metadata. *Journal of Library Metadata*, *11*(3-4), 155-165.

Jirotka, M., Procter, R., Hartswood, M., Slack, R., Simpson, A., Coopmans, C., Hinds, C., et al. (2005). Collaboration and Trust in Healthcare Innovation: The eDiaMoND CaseStudy. *Computer Supported Cooperative Work*, *14*, 369–398.

Karasti, H., Baker, K. S., & 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*, *2006*, 321–358.

Lave, J., & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press.

National Science Board. (2005). *Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century*. Washington, D.C.: National Science Foundation.

Palmer, C. L. (1996). Information work at the boundaries of science: Linking Library services to research practices. *Library Trends*, *45*(2), 165–191.

Pierce, S. J. (1999). Boundary crossing in research literatures as a means of interdisciplinary information transfer. *Journal of the American Society for Information Science*, *50*(3), 271–279.

Schiff, L. R., Van House, N. A., & Butler, M. H. (1997). Understanding complex information environments: a social analysis of watershed planning. *Proceedings of the second ACM international conference on Digital libraries*, DL '97 (pp. 161–168). New York, NY, USA: ACM.

Van House, N.A., Butler, M. H., & Schiff, L. R. (1998). Cooperative Knowledge Work and Practices of Trust: Sharing Environmental Planning Data Sets. *The ACM Conference on Computer Supported Cooperative Work* (pp. 335–343). Seattle, Washington.

Van House, N. A. (2002). Digital libraries and practices of trust: Networked biodiversity information. *Social Epistemology: A Journal of Knowledge, Culture and Policy*, *16*(1), 99.

Wallis, J. C., Milojevic, S., Borgman, C. L., & Sandoval, W. A. (2006). The Special Case of Scientific Data Sharing with Education. *The American Society for Information Science & Technology* (Vol. 43, pp. 169–181).

Weeks, P., & Packard, J. M. (1997). Acceptance of Scientific Management by Natural Resource Dependent Communities. *Conservation Biology*, *11*(1), 236–245.

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*, 5–16.

Zimmerman, A. S. (2008). New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data. *Science, Technology & Human Values*, *33*(5), 631–652.