The Role of Data Reuse in the Apprenticeship Process

Adam Kriesberg
University of Michigan
School of Information
105 S. State St
Ann Arbor, MI 48109
akriesbe@umich.edu

Rebecca D. Frank
University of Michigan
School of Information
105 S. State St
Ann Arbor, MI 48109
frankrd@umich.edu

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

Elizabeth Yakel
University of Michigan
School of Information
105 S. State St
Ann Arbor, MI 48109
yakel@umich.edu

ABSTRACT

The availability of research data through digital repositories has made data reuse a possibility in a growing number of fields. This paper reports on the results of interviews with 27 zoologists, 43 quantitative social scientists and 22 archaeologists. It examines how data reuse contributes to the apprenticeship process and aids students in becoming full members of scholarly disciplines. Specifically, it investigates how data reuse contributes to the processes by which novice researchers join academic communities of practice. We demonstrate how projects involving data reuse provide a unique opportunity for advisors to mentor novices through the process of creating knowledge. In these situations, senior researchers model general reuse practices and impart skills for their students to use in the future when selecting, evaluating, and analyzing data they did not collect. For novices, data reuse constitutes a form of legitimate peripheral participation, a way for them to enter the community of practice by analyzing data that has been previously collected and reflecting on others' methodologies. Our study findings indicate that reuse occurs across each target community studied. They also suggest how repositories can help foster a reuse culture by providing access to data and building trust in research communities.

Keywords

Data reuse, communities of practice, legitimate peripheral participation, cognitive apprenticeship, digital repositories, disciplinary repositories

INTRODUCTION

Increasingly, academic researchers need to acquire the ability to reuse data in order to create new knowledge. Given the importance of longitudinal data in some fields, the difficulties associated with collecting new data in others, and the fact that some types of data from the field need to be compared with existing materials, such as specimens, scholars from a variety of disciplines engage in data reuse. In this paper, we report on the results of 92 interviews with researchers in three academic communities: zoology, quantitative social science, and archaeology. We focus on the role data reuse plays in the apprenticeship process, helping novices become members of communities of practice through cognitive apprenticeship and legitimate peripheral participation. We argue that the concept of legitimate peripheral participation should be expanded to include a range of activities surrounding data reuse. These include activities traditionally within the scope of legitimate peripheral participation, such as engaging with other researchers, as well as reviewing the literature and analyzing the methodology behind a dataset in preparation for data reuse.

Our study is motivated by the following research questions:

- What role does data reuse play in the apprenticeship process to incorporate new members into academic communities of practice?
- How does data reuse extend our understanding of the intellectual, social, and structural mechanisms behind cognitive apprenticeship and legitimate peripheral participation?

LITERATURE REVIEW

Socialization into a community of practice has been described as a series of passages through which one must navigate (Van Maanen & Schein, 1979). Communities of practice provide a body of knowledge and skills as well as an infrastructure for the intellectual scaffolding necessary to make sense of that knowledge. This is often provided through mentoring wherein senior community members provide support, direction, and information to novice community members (Lave & Wenger, 1991; Van House, Butler, & Schiff, 1998). Lave and Wenger (1991) have identified one mechanism through which novices become

ASIST 2013, November 1-6, 2013, Montreal, Quebec, Canada.
Copyright notice continues right here.

members of a community of practice as legitimate peripheral participation. Legitimate peripheral participation “refers both to the development of knowledgeably skilled identities in practice and to the reproduction and transformation of communities of practice” (Lave & Wenger, 1991, p. 55). It places an emphasis on the context in which learning takes place and occurs when the novice assumes increasingly responsible roles in the research, beginning with observation of research activities and ending with directing the activities of others (Duguid, 2005; Lave & Wenger, 1991).

Students also become members of the community of practice by internalizing norms and methods of inquiry (Ben-Yehuda, 1986). Researchers have called this process cognitive apprenticeship, “a model of instruction that works to make thinking visible” (Collins, Brown, & Holum, 1991, p. 1). Cognitive apprenticeship goes beyond observation of the supervisor as role model and includes learning through doing and engaging as an actual practitioner in research (Anderson & Louis, 1994; Bragg, 1976; Brown, Collins, & Duguid, 1989; Clark & Corcoran, 1986; Collins et al., 1991). Collins, Brown, and Holum (1991) argue that cognitive apprenticeship is distinct from traditional models of apprenticeship because the mentor makes her thinking visible to the apprentice, the learning is situated in the workplace, and the goal is to help students generalize the skill. Still, although there is a skill-based component to cognitive apprenticeship, the goal is to structure the thinking process in the context of an advisor-advisee relationship. For example, the apprenticeship model of acculturation for archaeologists involves field work with the goals of collecting data under certain methodological norms, learning how to behave appropriately when working with culturally sensitive materials, and documenting evidence (Edgeworth, 1991; Pyburn, 2003). Mentoring in the area of quantitative social science focuses on research as a one-on-one activity between novices and mentors in which novices are mentored to identify and pursue their own research questions and theories, rather than work on research projects initiated by senior researchers (Anderson & Louis, 1994; Faniel, Kriesberg, & Yakel, 2012).

Much of the literature in the area of mentoring and apprenticeship in graduate school has focused on qualities and characteristics of the students and mentors as well as the relationship between them. Supervisors are often described as exemplars and mentors with a focus on the importance of integrity and ethics (Anderson, Oju, & Falkner, 2001; Gray & Jordan, 2012). The role of the mentor is often marked by a tension between the demands placed on them by their institution and the responsibility to support their students (Holligan, 2005).

Broadening the focus to include the apprentice as well as the mentor, researchers have noted that the quality of the mentoring relationship depends on both the supervisor and the student (Kam, 1997). Notably, students are more likely to be satisfied and to make good progress when they have a good relationship with their mentor, particularly when that mentor is an academic with senior status within their community of practice (Ives & Rowley, 2005). This apprenticeship relationship is marked by a duality of personal relations based on trust and social relations reinforced through contact. Researchers argue that this relationship is problematized by complications that arise due to the demands of research placed on both students and mentors (Denicolo, 2004; Hockey, 1996; Lee, 2008).

Researchers studying mentoring in the hard sciences have found that local setting, work group size, and discipline all affect the process of becoming a scientist (Louis, Holdsworth, Anderson, & Campbell, 2007). Scholars have also found that good mentoring in the sciences has positive benefits for productivity and self-efficacy, but not with commitment to a research career (Paglis, Green, & Bauer, 2006). These findings are particularly interesting as one of the purposes of apprenticeship is to help novice researchers to become members of a community of practice.

Novices engage in research activities with experts, and work with data in particular, through the application of standardized research methods that guide practices and behaviors (Star & Griesemer, 1989). Formal training within a community of practice leads to a familiarity with particular types of data (Zimmerman, 2007). This familiarity may then guide future research methods or data seeking behaviors. Additionally, novice social science researchers are influenced by more experienced members of their own community of practice when it comes to discovering, evaluating, and justifying their reuse of data (Faniel et al., 2012).

Using norms, academic disciplines organize life inside academic institutions, and the differing cultural values of the departments within those institutions affect the views of graduate students regarding the purposes and processes of research (Gieryn, 1983; Hackett, 1990). The process of acculturation and training via apprenticeship can be viewed as a means by which novices are brought into a community of practice. It is through interactions with mentors, through cognitive apprenticeship and legitimate peripheral participation in the research process, that novices are introduced to the cultures of their communities of practice.

While the literature does not focus on data reuse as a specific method of legitimate peripheral participation, we argue that data reuse is also a critical component of the process of acculturation for novice researchers into communities of practice because data reuse is predicated on understanding what constitutes data within the context of a discipline, and norms for its collection and interpretation.

METHODS

The Dissemination Information Packages for Information Reuse (DIPIR) Project is a three-year IMLS sponsored initiative studying three diverse disciplinary communities: zoology, archaeology, and quantitative social science

(Faniel & Yakel, 2011). The current paper draws upon interviews conducted with members of all three communities, primarily focusing on novice researchers. DIPIR Project research partners affiliated with repositories in each community (described below) helped facilitate access to potential participants from the three disciplines.

Site Descriptions and a Comparison of Communities

The University of Michigan Museum of Zoology (UMMZ) was founded in 1895 and houses some 15 million specimens across six divisions. It supports the research of scientists in a number of fields, and serves as a resource for students. In addition, UMMZ partners with other institutions to contribute data to a series of digital repositories focused on a particular group of animals (i.e. FishNet for data on fish specimens). These more centralized repositories are also access points for prospective data reusers, in addition to online museum catalogs, and visits to the institutions to perform additional analyses on physical specimens.

Open Context is an open access data publication platform for archaeological data. Founded in 2007 and maintained by the Alexandria Archive Institute, it is an emerging hub for both experienced and novice archaeologists looking to find primary data for reuse. The repository values contributions in the form of data and views data sharing as a form of publication, vital to advancing teaching and research in the archaeological community. It also supports open standards for data in its collection, with an eye towards facilitating reuse of its culturally valuable data across disciplines and repositories.

The Inter-University Consortium for Political and Social Research (ICPSR) was founded in 1962 (Vardigan & Whiteman, 2007) and is a leader in the field of social science data preservation, access, and curation. It holds more than 50,000 data files and serves diverse communities. Additionally, the consortium hosts a summer program, providing training in research methodology.

Some of the differences we observed in the interviews can be traced back to differences in academic cultures across the communities in which we worked. For example, the presence of ICPSR as an archive for social science data since 1962 has helped create research communities built upon strong traditions of data reuse. Some of the datasets available through ICPSR also encompass successive waves of data collection over decades. The long history of comparative and morphological analysis in zoology using museum collections constituted the foundations of a reuse culture, but the emergence of digital repositories that provide access to information across institutions has had a great effect on research, enabling projects with a broader scope and fostering more data sharing. Finally, archaeological repositories are in their infancy and have only recently begun to gain traction in the discipline, particularly in the UK. Scholars are turning to data reuse as the questions in archaeology are changing from a site-based focus to larger social, economic, and cultural trends. Increased availability of existing archaeological data will allow the community to extract as much value as possible from what is collected, and not allow data such as those contained within the "grey literature," to perpetually remain difficult to access.

Subject Recruitment

The team recruited subjects from each community using convenience and snowball sampling techniques. Novices from the quantitative social science community were recruited via targeted emails to participants in the 2011 ICPSR Summer Program. For the archaeology and zoology interviews, DIPIR project partners recruited a list of potential participants. In all three instances, additional participants were obtained using snowball sampling; we asked those we interviewed to recommend others who might be interested in participating in our study.

Data Collection

Team members conducted four sets of interviews across the three academic communities that are the focus of the DIPIR Project. We conducted semi-structured interviews with 42 quantitative social scientists (22 novices and 21 experts), 22 archaeologists, and 27 zoologists. While the interview protocol varied slightly in each set of interviews due to disciplinary specifics, each instrument asked respondents to reflect on how they discovered, evaluated, and analyzed data for reuse, and about their use of data repositories. Each interview lasted approximately 1 hour and respondents were paid $25 for their participation in the study. All interviews were audio recorded and transcribed for analysis.

In our interviews across communities, we defined novices as early career researchers such as graduate students and post-docs. While we conducted interviews with researchers in a variety of roles and with varied levels of experience, discussions of novice experiences form the basis of findings for the present study. Therefore data from senior scholars also appears in discussions related to the mentoring they received as novices or the mentoring they gave to novices.

Data Analysis

All interview transcripts were analyzed using the qualitative data analysis software package NVivo. After the team created an initial code set, two coders worked together on each set of transcripts from the three source communities. The initial development of the code set was based on themes from the interview protocol. Among the items coded for were mentions of the different dimensions of data reuse, including discovery, evaluation, methods, and data sharing. Transcripts were also coded for mentions of respondents interacting with peers or advisors or asking for help. Additional codes emerged from each set of interviews and were added as needed. For example, a code relating to the specific ethical challenges of archaeology was added to that set of interviews. After a series of paired work on the same transcript, the coders reached the following reliability ratings using Scott's Pi, a statistic measuring inter-rater reliability for coding textual data: 0.88 for novice social scientists, 0.77 for expert social scientists, 0.73 for archaeologists, and 0.74 for zoologists.

FINDINGS

Interviewees related experiences which mapped well to the themes we explored in the literature. Specifically, we observed that the intellectual process of working through data originally produced by another and the development of skills around data reuse – discovery, evaluation, and analysis of data – were essential components of participants' maturation as researchers. We argue that learning how to reuse data, and create knowledge using data that interviewees did not collect, provides insight into the norms in their disciplinary communities of practice. Participants in the three disciplines under study followed parallel paths in learning how to reuse data and how to participate in their disciplinary communities. In addition to the similarities observed across our three groups of interviews, discipline-specific differences emerged.

In all three disciplines under study, students and advisors worked together to identify relevant data for reuse and to gain access to that data in order to conduct analyses. This reflects what we learned from the literature. Namely, that the relationship between graduate student and advisor is fundamental to the graduate school experience (Ben-Yehuda, 1986). Once the desired data was acquired, novice researchers engaged in a variety of analytic techniques to accomplish dual goals: completion of a given project and the more ambitious long term effort to become a member of a chosen research community by understanding disciplinary norms and what constitutes a research contribution.

The rest of this section is presented in four parts. First, we briefly compare the reuse cultures across our three disciplines of interest to help contextualize our study. Then we present findings about the mentor/mentee relationship formed between graduate students and advisors, highlighting the importance of data competency in this process. Third, we argue that data reuse supports cognitive apprenticeship as a way for novice researchers to learn about data sharing culture and norms in their field. Finally, we make the case that learning to reuse data is a form of legitimate peripheral participation, used by novice researchers to gain entry into their chosen community of practice.

The Role of Data Reuse in Quantitative Social Science, Archaeology, and Zoology

Across our interviews, we noted the differences in the role of data reuse. For instance, researchers reused data to compare with data they collected, engage in longitudinal analysis, or test their hypotheses with a larger, nationally representative sample. Beyond the need to reuse data to pursue larger scale research projects, we found differences in research methods and data sharing cultures, as well as generational shifts in the drive to reuse data within each discipline.

In archaeology, norms around data sharing and reuse are just emerging and not fully integrated into the community of practice. Archaeological publications do not consistently include the data in addition to the article interpreting the data. As archaeologist CCU17 described, one primary challenge for data reuse in his field was discovery and access. Not only was there a significant lag between data collection and publication, there was also a lack of centralized access. He hoped that a new network of digital repositories could attract submissions to help alleviate this issue.

There's often such a lag between fieldwork and publication. Or you might not know where the publications have been made, if they have been made, for a certain project…you've got monographs that come out five years after a 10-year project ends. It's kind of a fairly big lag there, and so it would be nice to have some ideas of where people are working. And I guess, we could also include commercial work, which should be dead useful because that grey literature is often pretty ignored, I think, in research. So, it would be useful to have everything pulled into one place (CCU17).

The access challenges faced by archaeologists differed from the active data reuse culture described by social scientists. The influence of well established repositories like ICPSR, has had a normalizing effect. Data reuse is not only possible, but also an accepted and valued part of the research community. CBU09 recognized this during her time in graduate school, observing the research and publishing habits of her professors.

I think the gold standard is collecting your own [data], but that's not always an option. I would say that a majority of the government professors have not collected their own data for let's say their last couple of articles. I worked with one professor very closely and he had four articles that he was working on and one of them was original data (CBU09).

While data reuse was embraced in the quantitative social sciences, the younger generation of zoologists described a generational shift of opinion on the value of digital repositories and the research contributions that can be made. Even though reference collections in museums have been an accepted part of the zoological practice for years, the availability of digital data has prompted discussion about the value of data reuse and its place in the disciplinary community.

It's a generational thing…and so there's a growing pain that the field is experiencing right now…folks who haven't used these tools because maybe they got their PhD 30 years ago and these databases weren't important in their research…and so there's some friction in there. Younger folks who understand the importance of these databases *and synthetic projects, and then folks who would never use them because they weren't around…but I think more and more that reticence to use data that aren't your own, I think that that's subsiding and folks are realizing that this is the way that systematics is going to go (CAU02).*

CAU02 saw himself as a member of the new generation of zoologists. He felt that he needed to demonstrate, through his research, the value of data reuse in his field. While he went on to describe additional tensions between his peers and some senior members of the field, he believed that an increased acceptance of data reuse could advance the field and reduce redundant studies through the sharing of data between colleagues.

Despite their differences, interviewees in each of our communities of focus expressed positive feelings towards data reuse and optimism about its future. However, these disciplinary particulars underlie the variation across the rest of our results. While novices across disciplines engaged in data reuse as legitimate peripheral participation and learned in the context of their cognitive apprenticeships, the particular form of their reuse was driven by discipline. Generally, reuse is an established part of the social sciences; a reality reflected in what we heard from respondents about their experiences joining communities of practice. In zoology and archaeology, data reuse takes different forms, in part because of generational tension and a less clearly defined culture around digital data.

Cognitive Apprenticeships

Lessons learned from an advisor through a cognitive apprenticeship process in graduate school helped novice researchers understand community norms and the steps necessary to do research in a given discipline. In our source communities, data reuse was a central component of the mentorship process. CCU19, an archaeologist, was a senior scholar and served as advisor to a student on a recent project involving analysis of an older dataset, a situation which provided an opportunity for mentorship.

I'm mentoring a student in a research project right now. And what we did was we went to a site which made available information about Mayan site locations…I had her download the data and start to evaluate it in terms of how updated it was and what we would need to do in order to start gathering information and bring it up to date. So in that case, we were basically taking a partially updated dataset…using that as a foundation for this student's project, which then will be shared back with other people (CCU19).

This example presents the mentor relationship from the perspective of the faculty member. He guided a student through the process of evaluating a dataset for reuse, demonstrating in the context of a real project what is required to constitute a reusable dataset. Furthermore, by assuring that the updated dataset will be shared back with the community, he instilled the student with values that will reinforce data sharing in the field.

For social scientist CBU09, the process of cognitive apprenticeship involved placing trust in faculty members to help guide her analysis. Given that CBU09 was in the process of applying to graduate schools and not yet enrolled in a program, her ability to conduct statistical analyses of datasets was limited. While she began the process herself, the directed guidance of faculty members provided meaningful next steps towards constructing a complete analysis.

I am new to this. And there's a lot of things that are counter-intuitive, that you just need someone to tell you, 'No, it's this way...' (CBU09).

CBU09 was in a position where she was not able to accomplish data reuse on her own. Without the input of a mentor to keep her on track, she would have been lost. Recognizing this, she sought the advice of senior scholars to show her ways to engage with the data.

Across our three communities, novice researchers began projects involving data reuse with little fanfare, as a normal and expected part of their graduate school research experience. For CAU24, his mentor assigned him data for reuse.

My advisor has been working with another professor in order to describe this endangered species of cave snails in Illinois…she had started working on it and I was on a research assistantship, so it ended up where I ended up working on it. So, really, it was just one of those things where…it was just the project I was assigned (CAU24).

For many graduate students in Zoology (CAU24's field), it was a challenge to collect their own data. However, there were also expectations within the discipline that data reuse would be necessary to compare museum specimens with those they were able to collect in the field. For these reasons, it makes sense that CAU24 would be assigned a data reuse project in his early years in graduate school so he could begin to engage in common research activities.

In the social sciences, graduate students often reuse datasets for their dissertation work. CBU10 described the discovery of data for her dissertation as a process initiated by an article recommended by a professor.

One of my professors asked me to include a couple of articles as a part of my literature review that pertained to my dissertation topic. And the article that really fascinated me…referenced Add Health where they were looking at tracking immigrants for over a period of time in the United States. It was through that article really that I found about Add Health. After a couple of my professors indicated to me that my approach to my research would take too long…they asked me to look for existing datasets. And so, that's how I came across Add Health (CBU10).

The second part of the example above shows another mentorship moment. CBU10's original data collection plan was too ambitious for a dissertation and was not going to be possible to complete. Remembering Add Health, she worked with her committee to design a project including reuse of the dataset that would constitute a meaningful addition to her field.

In these examples, we see data reuse as a focal point of learning during a cognitive apprenticeship. Through reusing data, students grapple with the conceptual issues of data selection and integration, within the context of relationships with advisors. As they learn how to make contributions of new knowledge in their respective fields, they come to understand that data reuse is a viable option in situations where original data collection is not possible or the goals of the project necessitate reuse.

Data Reuse as Pathway to Disciplinary Enculturation: Ethics, Evidential norms, and Disciplinary Culture

Beyond data selection and analysis, data reuse was also a pathway for cognitive apprenticeship in other aspects of disciplinary culture, such as ethics, understanding the norms for evidence, and how different disciplines approach research issues.

Mentors can also help pass along ethical norms, such as those promoting data sharing. Although CCU09 was a senior faculty member, she recalled the role her advisor played in instilling the value of data sharing. Although she had obtained her PhD years before, she vividly recalled lessons learned from her advisor early in her graduate school career.

He said, 'Even though you're going to summarize it and do whatever you're going to do with it and do statistics, et cetera.' He said, 'Somebody will come a long later and say you're wrong.' But he said, 'Your data is going to be the most important contribution you make.' And so he taught me to always, as an appendix, include all of my data. So then anybody could come along and redo what I did and see if I was right or if I was wrong or use it for something else…and so that was really sort of ingrained in me that that was a necessary thing because otherwise I was doing something bad (CCU09).

The above example also highlights a theme we did not often hear in our interviews, that of an advisor defining the difference between acceptable and unacceptable behavior, and explicitly warning a student about the potential for her results to be challenged after publication. When it came time to advise a new generation of students, CCU09 passed along the values that her mentor had taught her years before.

Another integral part of the relationship between faculty advisor and graduate student in the cognitive apprenticeship process is teaching a new scholar how to make and support claims in research when reusing data. What may seem interesting to a scholar in training may elicit a different reaction from a more senior faculty member. CBU11, a doctoral student in Political Science, experienced this when analyzing a dataset examining political polarization in the United States.

I [had] some sentence saying, 'Through the last decade, polarization has...been increasing steadily,' and [my advisor] said, 'Oh, not really. There were a few years here and there...' So I was able to go and really find out where it was steadily increasing versus when it was decreasing at all (CBU11).

Her advisor brought her back to the data for a closer examination. Thus, CBU11 was gently directed to scrutinize datasets more before making claims. In addition to the specifics of this situation, she learned what level of precision was required for claims in her chosen field, a lesson in knowledge production from her advisor.

As a graduate student in political science, CBU18 was working on a project trying to incorporate aspects of his home discipline as well as psychology. When talking about how he arrived at a final determination of whether to reuse a particular dataset, he pointed to his mentor as a key factor in the decision-making process.

…it was advice from my advisor. So I said, 'Well, what I'm doing is a lot more psychological.' The dataset I'm using is a lot more political science. How do we merge the two, such that, I can at least attempt to study what I'm trying to study with the number of subjects that would be acceptable in political science to say, 'Alright. I can run the type of data analyses that I need to run.' So it was mainly mentorship and my mentor advising me (CBU18).

In discussion with his advisor, CBU18 worked to apply his psychology-based approach in a way that would be acceptable in political science. As he described it, mentorship played a key role in his understanding and deciding to move ahead with the project.

When describing their data reuse experiences, our respondents discussed related disciplinary norms and ethical considerations beyond the scope of one specific project. Through the reuse of data, they also began to internalize the disciplinary culture of their communities of practice.

Data Reuse as Legitimate Peripheral Participation

In learning how to create knowledge that will be accepted by members of a given academic discipline, novice researchers in our three target communities related their experiences in reusing data. These activities demonstrated that through data reuse, they continued the process of joining communities of practice. For CAU17, the process of identifying suitable datasets for reuse was informed by his observations of colleagues.

If I'm putting a bunch of different sources of data together, I want to make sure that they were acquired in roughly the same way. So, for example, a lot of folks that I've worked *with, and I have done this a little bit too, is if I am creating one of these models using a bird species…I might develop the model using just museum specimens (CAU17).*

This example shows a novice researcher making a decision about which types of data to include in an analysis based on observed behavior of other researchers with whom he worked. Rather than making the decision in isolation, he looked to more experienced members of the community for a model of how to construct a suitable dataset for reuse.

For political science doctoral student CBU04, models for data reuse came from the literature. In this particular example, CBU04 used the data producer's publications in his evaluation of a dataset's suitability for his research. He talked about employing them as a means to look over the data producer's shoulder to understand the data collection process. CBU04 also looked to the literature for validation that a given measure had been successfully reused by other researchers as well as the codebook to understand the decisions behind data collection.

As long as I can trust how they measured things. If they followed standard sampling procedures, if this is like a survey research or something of that nature, or if they're just very, very explicit with this is the way I'm measuring this concept, that's very important to me…If this measure has been used by other authors, I tend to use that measure just because I might not understand why it's being used, but if it's being used, there's probably good reason (CBU04).

In addition to the trust between advisors and graduate students discussed earlier, the example above demonstrates how the literature and codebooks serve as a means of providing peripheral participation in the data collection process for novices. This type of engagement with additional forms of data documentation is a form of legitimate peripheral participation. By studying the literature and codebooks, the novice learns what types of data can be reused, and how to frame reuse in her publications.

When zoologist CAU27 began his thesis project, he realized that travelling to South America to collect specimens relevant to his interests was not realistic. Instead, he worked with his advisor to build on existing specimens and measurements to design a feasible project.

My advisor did a project during his postdoc and I'm using some of his measurements already in my research…as a stepping stone to learn my, or to develop my methods and integrate his previous methods into what I'm doing now, so I'm using some of his data as well (CAU27).

The legitimate peripheral participation in this case stems from the fact that the student is continuing work started by the advisor, including existing projects or data that the advisor and his colleagues might have collected but not analyzed to the fullest possible extent. Through close collaboration and affiliation with an advisor's project, CAU27 could pursue his own research questions outside those of his advisor using some of the same data.

Other participants implicitly understood their time in graduate school as a process of joining their discipline's community of practice. Criminal Justice doctoral student CBU05 reflected on her transition to graduate level work and her changing relationship to potentially reusable data. While she had some experience with data reuse as an undergraduate, the environment of a graduate program was pushing her to engage in more advanced data reuse practices. She acknowledged the changing expectations that reflect her movement towards being a full member of the community of practice.

Probably the amount of time that you do have to spend with the codebook especially with larger data sets…there's a lot more work that you have to do, not necessarily interactive work with people in interviewing, but you do have to spend a lot of time with the codebooks… [understanding] where those numbers are coming from and what they mean and hopefully to be able to trust them and make sure they're measuring what you're wanting to measure (CBU05).

In comparison to her earlier work, more time was required to evaluate and understand a codebook prior to data reuse. Through working with the codebook, the student learned about data collection techniques. While CBU05's undergraduate experience was a useful introduction to the field of Criminal Justice in that it sparked enough interest to prompt the pursuit of a graduate degree, it was only the first step in the process of joining this community of practice. As a graduate student advances in her studies, the student moves in from the periphery where she observes to full participation and reflection upon the experience gained through cognitive apprenticeship.

Our data indicated that respondents were conscious of their positions in their fields, and engaged in acts of legitimate peripheral participation while they learned how to be members of their chosen communities of practice. While legitimate peripheral participation is usually characterized as direct interactions with advisors, senior community members and peers, we also found that additional activities including observation of data reuse in the literature and the construction and critique of codebooks reflect the same concept. This extension of legitimate peripheral participation is especially useful when considering that our participants searched widely for models of how to act when engaging in reuse. When trying to reuse data, referring to the literature becomes more than simply reading papers; it is legitimate peripheral participation because novice researchers observe in the literature behavior that they then replicate in their own projects.

DISCUSSION

We drew upon interviews with 92 archaeologists, quantitative social scientists, and zoologists to examine the role of data reuse in the socialization process for novice researchers. Our findings indicated that learning how to find, evaluate, and analyze data for reuse is part of the training process understood as a cognitive apprenticeship, and that engaging with the data is itself an act of legitimate peripheral participation. We also saw disciplinary differences in the extent to which data reuse norms and values are shared within each of the three communities under study. Moreover, our findings indicate that data reuse occurs and contributes to the socialization of novices into communities of practice in multiple ways.

The literature on cognitive apprenticeship stresses the importance of the relationship between student and advisor. Mutual trust is needed for any successful partnership; in graduate school this relationship is critical because advisors are the gatekeepers for students, guiding them into communities of practice (Kam, 1997). Our findings suggest that data reuse is an opportunity for novice researchers and advisors to put concepts around knowledge creation into practice. Given advisors' positions, they guide new community members through the data reuse process, using these types of projects to not only to demonstrate specific techniques for data selection, evaluation, and analysis, but also to impart broader lessons about disciplinary ethics, norms and what constitutes a research contribution.

Across our three disciplines, we argue that data reuse is an important part of the socialization into communities of practice. While the goals and techniques of reuse vary in each discipline and across participants' projects, we found that novice researchers engage in data reuse, and that some of these activities take the form of legitimate peripheral participation. When novice researchers engaged in reuse projects, they consulted advisors and other senior researchers throughout the process. Beginning as outsiders, they are expected to move toward fuller participation on their way to joining communities of practice. They looked to the literature to understand disciplinary research processes and to find models of successful reuse. They also analyzed codebooks and other documentation about datasets to consider their appropriateness for reuse. Through these activities, the novice researchers we interviewed learned how to reuse data and make contributions to their fields. Our findings coincide with and expand the idea of legitimate peripheral participation as described by Lave & Wenger (1991) by expanding the activities around which learning takes place for novice researchers to include data reuse.

Similar to Duguid (2005), we seek to extend the scope of legitimate peripheral participation as a concept. Our findings suggest that data reuse has a role in the apprenticeship process and outlines a wider range of activities than previously analyzed. Our participants described a number of ways in which they engaged in legitimate peripheral participation through data reuse. They learned from advisors and senior researchers as cognitive apprentices, they collaborated with peers to learn about data reuse practices, and they looked to the literature for insight into the creation and analysis of datasets and actual ways it was reused. Codebooks also allowed novices to understand how to reuse data. These findings extend the situations in which legitimate peripheral participation has been found previously and demonstrate that the process of joining a community of practice involves not only original data collection activities, but also encompasses data reuse.

Implications for Data Repositories

Because novice researchers in archaeology, zoology, and quantitative social science reported using digital repositories to access data for reuse, our findings have implications for these organizations. Earlier work has demonstrated that repositories have the ability to shape norms for reuse in fields by working with their designated communities and responding to user feedback (Daniels, Faniel, Fear, & Yakel, 2012). Our findings reinforce this idea; the novice researchers we spoke to across disciplines were very aware of the community and culture surrounding the repositories they used. Yakel et al. (2013) found that repositories build trust in part by getting users to recognize their actions and develop trust in these institutions. Our findings on the reuse of data by novices during their period of socialization into their chosen community of practice speak to the affordances of easily-available and well curated digital data. When repositories develop strong scaffolding, such as well-written codebooks and links to data citations, novice researchers can readily see how data production and analysis match disciplinary norms and how data can be appropriately reused to answer their research questions. Yet, this also places an added burden on repositories if their data is to support cognitive apprenticeship and legitimate peripheral participation.

CONCLUSION

The availability of data for reuse and the high costs and relative difficulty graduate students encounter in collecting their own data creates an environment in which reuse is increasingly common. We interviewed 92 researchers in archaeology, zoology, and the quantitative social sciences who engaged in reuse, seeking to understand how these activities were situated in the context of novice researchers' cognitive apprenticeships, and in their efforts to join communities of practice. We found that data reuse functioned as a form of legitimate peripheral participation, a way for novices to conduct research in the community of practice while still learning how to be fully functioning members of their communities. Our data show that legitimate peripheral participation around data reuse takes the form of interactions with senior researchers, peers, and the literature. The consistency of these findings across three disparate disciplines suggests that the ability to reuse data is an important skill for researchers across the academy, one that can be supported through the expanded efforts of research communities and repositories to increase access to reusable data.

We hope this study encourages further inquiry about data reuse in the education and training of novice researchers. While results from only three fields are presented here, we think that future research into other disciplines (e.g. lab sciences, additional humanities disciplines) may involve a new set of reuse activities which support cognitive apprenticeships and legitimate peripheral participation in different ways from those presented in this paper. Through further research from the user and repository perspectives, more insights will emerge about how to best support the apprenticeship process through data reuse across the academy.

ACKNOWLEDGMENTS

The DIPIR Project 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.” We also thank other members of the DIPIR team who participated in data collection and analysis efforts, including Jessica Schaengold, Julianna Barrera-Gomez, Morgan Daniels, and Ayoung Yoon.

REFERENCES

Anderson, M. S., & Louis, K. S. (1994). The Graduate Student Experience and Subscription to the Norms of Science. *Research in Higher Education*, *35*(3), 273–299. doi:10.1007/BF02496825

Anderson, M. S., Oju, E. C., & Falkner, T. M. R. (2001). Help from Faculty: Findings from the Acadia Institute Graduate Education Study. *Science and Engineering Ethics*, *7*(4), 487–503. doi:10.1007/s11948-001-0006-x

Ben-Yehuda, N. (1986). Deviance in Science Towards the Criminology of Science. *British Journal of Criminology*, *26*(1), 1–27.

Bragg, A. K. (1976). *The Socialization Process in Higher Education* (No. ERIC/Higher Education Research Report No. 7.). Washington, D.C.: American Association for Higher Education. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED132909

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated Cognition and the Culture of Learning. *Educational Researcher*, *18*(1), 32–42.

Clark, S. M., & Corcoran, M. (1986). Perspectives on the Professional Socialization of Women Faculty: A Case of Accumulative Disadvantage? *The Journal of Higher Education*, *57*(1), 20–43. doi:10.2307/1981464

Collins, A., Brown, J. S., & Holum, A. (1991). Cognitive Apprenticeship: Making Things Visible. *American Educator: The Professional Journal of the American Federation of Teachers*, *15*(3), 6–11,38–46.

Daniels, M., Faniel, I., Fear, K., & Yakel, E. (2012). Managing Fixity and Fluidity in Data Repositories. In *Proceedings of the 2012 iConference* (pp. 279–286). Presented at the iConference 2012, Toronto, Canada: ACM. doi:10.1145/2132176.2132212

Denicolo, P. (2004). Doctoral Supervision of Colleagues: Peeling Off the Veneer of Satisfaction and Competence. *Studies in Higher Education*, *29*(6), 693–707. doi:10.1080/0307507042000287203

Duguid, P. (2005). “The Art of Knowing”: Social and Tacit Dimensions of Knowledge and the Limits of the Community of Practice. *The Information Society*, *21*(2), 109–118. doi:10.1080/01972240590925311

Edgeworth, M. (1991). *The Act of Discovery: An Ethnography of the Subject-Object Relation in Archaeological Practice*. Durham University, Durham, UK. Retrieved from http://etheses.dur.ac.uk/1481/

Faniel, I. M., Kriesberg, A., & Yakel, E. (2012). Data Reuse and Sensemaking Among Novice Social Scientists. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1–10. doi:10.1002/meet.14504901068

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

Gieryn, T. F. (1983). Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists. *American Sociological Review*, *48*(6), 781–795. doi:10.2307/2095325

Gray, P. W., & Jordan, S. R. (2012). Supervisors and Academic Integrity: Supervisors as Exemplars and Mentors. *Journal of Academic Ethics*, *10*(4), 299–311. doi:10.1007/s10805-012-9155-6

Hackett, E. J. (1990). Science as a Vocation in the 1990s: The Changing Organizational Culture of Academic Science. *The Journal of Higher Education*, *61*(3), 241–279. doi:10.2307/1982130

Hockey, J. (1996). A Contractual Solution to Problems in the Supervision of Phd Degrees in the UK. *Studies in Higher Education*, *21*(3), 359–371. doi:10.1080/03075079612331381271

Holligan, C. (2005). Fact and Fiction: A Case History of Doctoral Supervision. *Educational Research*, *47*(3), 267–278. doi:10.1080/00131880500287179

Ives, G., & Rowley, G. (2005). Supervisor Selection or Allocation and Continuity of Supervision: Ph.d. Students’ Progress and Outcomes. *Studies in Higher Education*, *30*(5), 535–555. doi:10.1080/03075070500249161

Kam, B. H. (1997). Style and Quality in Research Supervision: The Supervisor Dependency Factor. *Higher Education*, *34*(1), 81–103. doi:10.1023/A:1002946922952

Lave, J., & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge [England]; New York: Cambridge University Press.

Lee, A. (2008). How Are Doctoral Students Supervised? Concepts of Doctoral Research Supervision. *Studies in Higher Education*, *33*(3), 267–281. doi:10.1080/03075070802049202

Louis, K. S., Holdsworth, J. M., Anderson, M. S., & Campbell, E. G. (2007). Becoming a Scientist: The Effects of Work-Group Size and Organizational Climate. *The Journal of Higher Education*, *78*(3), 311–336.

Paglis, L. L., Green, S. G., & Bauer, T. N. (2006). Does Adviser Mentoring Add Value? A Longitudinal Study of Mentoring and Doctoral Student Outcomes. *Research in Higher Education*, *47*(4), 451–476. doi:10.1007/s11162-005-9003-2

Pyburn, K. A. (2003). What Are We Really Teaching in Archaeological Field Schools? In L. J. Zimmerman, K. D. Vitelli, & J. Hollowell-Zimmer (Eds.), *Ethical Issues in Archaeology* (pp. 213–223). Walnut Creek, CA: AltaMira Press.

Star, S. L., & Griesemer, J. R. (1989). Institutional Ecology, `Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, *19*(3), 387–420. doi:10.1177/030631289019003001

Van House, N. A., Butler, M. H., & Schiff, L. R. (1998). Cooperative Knowledge Work and Practices of Trust: Sharing Environmental Planning Data Sets. In *Proceedings of the 1998 ACM Conference On Computer Supported Cooperative Work* (pp. 335–343). Presented at the CSCW '98, Seattle, Washington: ACM. doi:10.1145/289444.289508

Van Maanen, J., & Schein, E. H. (1979). Toward a Theory of Organizational Socialization. In B. Staw (Ed.), *Research in Organizational Behavior* (Vol. 1, pp. 209–264). Greenwich, CT: JAI Press, Inc.

Vardigan, M., & Whiteman, C. (2007). ICPSR Meets OAIS: Applying the OAIS Reference Model to the Social Science Archive Context. *Archival Science*, *7*(1), 73–87. doi:10.1007/s10502-006-9037-z

Yakel, E., Faniel, I. M., Kriesberg, A., & Yoon, A. (2013). Trust in Digital Repositories. *International Journal of Digital Curation*. doi:10.2218/ijdc.v8i1.251

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. doi:10.1007/s00799-007-0015-8